DOCUMENT RESUME

ED 066 893 EM 010 196

AUTHOR	Bergman, Brian A.; Siegel, Arthur I.
TITLE	Training Evaluation and Student Achievement Measurement: A Review of the Literature.
INSTITUTION	Applied Psychological Services, Inc., Wayne, Pa.
SPONS AGENCY	Air Force Human Resources Lab., Lowry AFB, Colo.
REPORT NO	AFHRL-TR-72-3
PUB DATE	Jan 72
NOTE	67p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Achievement; Branching; Computer Assisted Instruction; Confidence Testing; Cost Effectiveness; Criterion Referenced Tests; Curriculum Development; *Evaluation Methods; Instructional Technology; Learning Modalities; *Literature Reviews; *Measurement Techniques; Military Training; Motivation; Statistical Analysis; Systems Approach; Testing; *Training Techniques

ABSTRACT

Training evaluation and student achievement measurement literature is reviewed with primary emphasis placed on studies reported in the last 10 years. Recent trends in training evaluation and student achievement measurement are presented. Factors relating to this topic, such as statistical methods, course development methods, training techniques, learning styles, motivation, and moderator variables are also included. Where new methods of training evaluation and student achievement measurement appear in the literature, detailed presentations are given. Among these procedures were cost-effectiveness or cost-benefit analysis, criterion-referenced testing, sequential testing, confidence testing, convergent and discriminant validity, and computer-assisted branched testing. Conclusions are that systematic approaches to evaluation and course development are receiving more and more attention. Most systems begin with a job analysis in order to derive a list of behaviorally-oriented job requirements from which training objectives can be formulated. The new techniques in evaluation and measurement have resulted from attempts to determine whether training objectives have been realized. (Author/JK)

ED 066893

AIR FORCE HUMAN RESOURCES LABORATORY

AFHRL-TR-72-3

TRAINING EVALUATION AND STUDENT ACHIEVEMENT MEASUREMENT: A REVIEW OF THE LITERATURE

By

Brian A. Bergman
Arthur I. Siegel
Applied Psychological Services, Inc.
Wayne, Pennsylvania

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado

January 1972

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND

BROOKS AIR FORCE BASE, TEXAS

EM 010 196

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

ED 066893

AFHRL-TR-72-3

January 1972

TRAINING EVALUATION AND STUDENT ACHIEVEMENT MEASUREMENT: A REVIEW OF THE LITERATURE

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

By

Brian A. Bergman
Arthur I. Siegel
Applied Psychological Services, Inc.
Wayne, Pennsylvania

Approved for public release; distribution unlimited.

TECHNICAL TRAINING DIVISION
AIR FORCE HUMAN RESOURCES LABORATORY
AIR FORCE SYSTEMS COMMAND
Lowry Air Force Base, Colorado

FOREWORD

This research was completed under Project 1121, Technical Training Development; Task 112103, Evaluating Individual Proficiency and Technical Training Programs. Dr. Marty R. Rockway was the Project Scientist, and Capt Wayne S. Sellman was the Task Scientist.

The research was carried out under the provisions of Contract F41609-71-C-0025 by Applied Psychological Services, Inc., Wayne, Pennsylvania. Project Monitor was Capt Wayne S. Sellman. Other reports prepared under this contract include AFHRL-TR-72-15, *Handbook for the Evaluation of Technical Training Courses and Students*, and AFHRL-TR-72-17, *A Survey of Student Measurement and Course Evaluation Procedures within the Air Training Command*

This report has been reviewed and is approved.

George K. Patterson, Colonel, USAF
Commander

ABSTRACT

The current training evaluation and student measurement literature is reviewed. The emphasis is on studies which have been reported in the last ten years, although earlier studies which have impacted heavily on recent trends are also included. Because of the obvious interaction between both training evaluation and student measurement, on the one hand, and such topics as statistical methods, methods for course development, training methods, learning styles, motivation, and moderator variables, on the other hand, these and similar considerations are also included.

SUMMARY

Bergman, B.A., & Siegel, A.I. *Training evaluation and student achievement measurement: A review of the literature.* AFHRL-TR-72-3. Lowry AFB, Colo.: Technical Training Division, Air Force Human Resources Laboratory, January 1972.

Problem

The purpose of this paper is to review the training evaluation and student achievement measurement literature with primary emphasis being placed on studies reported in the last ten years.

Approach

Recent trends in training evaluation and student achievement measurement are presented. Because of the obvious interaction between both training evaluation and student measurement, on the one hand, and such topics as statistical methods, course development methods, training techniques, learning styles, motivation, and moderator variables, on the other hand, these and similar considerations are also included.

Results

Where new methods of training evaluation and student achievement measurement appeared in the literature, detailed presentations were given. Among these procedures were cost-effectiveness or cost-benefit analysis, criterion-referenced testing, sequential testing, confidence testing, convergent and discriminant validity, and computer assisted branched testing.

Conclusions

Systematic approaches to evaluation and course development are receiving more and more attention. Most systems begin with a job analysis in order to derive a list of behaviorally oriented job requirements from which training objectives can be formulated. The new techniques in evaluation and measurement have resulted from attempts to determine whether training objectives have been realized.

This summary was prepared by Wayne S. Sellman, Technical Training Division, Air Force Human Resources Laboratory.

TABLE OF CONTENTS

	Page
I. Introduction	1
Sources Searched	1
Training Evaluation and Student Achievement Measurement	2
II. Dimensions of Evaluation	2
Roles, Uses, and Characteristics of Evaluation	2
Specification of Objectives	4
Systematic Approaches to Course Development	4
Measures and Methods of Evaluation	6
Problems of Evaluation	8
Summary	9
III. Quantitative Methods and Dependent Measures	9
Characteristics of Dependent Variables	9
Test Construction	10
Hierarchical and Sequential Testing	11
Criterion- and Norm-Referenced Testing	12
Ratings	13
Cost Effectiveness	14
Gain Scores and Final Examination Grades	16
Confidence Testing and Partial Knowledge	16
Characteristics of Material to be Learned	17
Composition Scoring	18
Testing	18
Statistical Methods	18
Partial Correlation and Part Correlation	18
Factor Analysis	19
Canonical Correlation	19
Moderator Variables	19
Convergent and Discriminant Validity	19
Internal and External Validity	20
Scaling Techniques	20
Signal Detection	21
Psychophysics	21
Summary	22
IV. Learning Styles and Moderator Variables	22
Scope of the Problem	22
Motivation and Types of Intelligence	22
Race and Aptitude as Moderator Variables	24
Age and Sex as Moderators	25
Cross-National Evaluation	25
Summary	25

Table of Contents *(Continued)*

	Page
V. Current Trends	25
Trends	25
Predictive Evaluation	26
Sensitivity Training	27
Programmed Instruction	28
Computer Assisted Instruction (CAI) and Testing	29
Applications of Programmed Instruction	30
Programmer Characteristics	31
Television Instruction	31
Basic Education	32
Training Devices	32
Instructor Evaluation	33
Military Research	33
Summary	35
VI. Comparative Evaluation	35
Comparative Studies of Subjects Within Average or Higher Aptitude Ranges	35
Comparative Studies of Low-Aptitude Subjects	40
Summary	41
VII. Discussion	42
References	44

TRAINING EVALUATION AND STUDENT ACHIEVEMENT MEASUREMENT: A REVIEW OF THE LITERATURE

I. INTRODUCTION

Methods and procedures for evaluating training courses and student achievement have been slowly evolving and assuming increased stature within any training program developmental paradigm which aims to be at all complete. This increased emphasis on training evaluation and student measurement is due, in part, to the increased realization that there can be no training system without quality control. Training in this sense is viewed as a process (analogous to a chemical or manufacturing process) in which raw material (students) is converted from one form to another (skilled craftsmen). Within such a construct, there must be a quality control stage; training evaluation and student measurement represent the quality control stage in the training process.

This report selectively reviews the current literature related to training evaluation and student achievement measurement. The review period extends over the 20 years preceding 1970, although the emphasis is not evenly apportioned throughout the entire span. The first ten years of the period are only briefly covered. Advances of the last decade indicate that, except for historical perspective, the 1950 to 1960 time frame should be treated rather lightly in a review such as this. Air Force flight equipment of the Korean War and immediate post-Korean War era is today looked upon as vintage equipment. Ten years ago, the digital computer, systems thinking, and programmed instruction were in their virtual infancy; and computer assisted training, T-group training, and behavior modification were all things of the future. Accordingly, the first decade of the review period has received only modest emphasis.

The heavier emphasis in this review is the recent ten years, with the last five being most thoroughly covered. The goal was to examine the subject matter areas but, most importantly, to determine for future reference, the answers to the questions "what is new in training evaluation?" and "what is new in student achievement measurement?" With these principal goals, placement of heaviest emphasis on the most contemporary time period seems clearly indicated.

Sources Searched

In order to identify relevant literature, the following sources were searched: *Psychological Abstracts, Technical Abstract Bulletins* of the Defense Documentation Center, and the *U. S. Government Research and Development Reports*, published by the Department of Commerce.

The *Psychological Abstracts* were reviewed from Number 1 of the 1966 volume through Number 4 of the 1971 volume, thus affording entry to the literature of the 1965-1970 period. The topics covered were Education and Training in the General section; Testing in the Methodology and Research Technology section; Testing, Counseling and Guidance, Teachers and Teacher Training, School Learning and Achievement in the Educational Psychology section; and Vocational Choice and Guidance, Selection, and Placement, and Training, in the Personnel and Industrial Psychology section.

The *Technical Abstract Bulletins* were reviewed from Number 1 of the 1966 index volume to Number 24 of the 1970 volume. The topics searched in these index volumes were Evaluation, Performance, Personnel, and Testing.

The *U. S. Government Research and Development Reports* reviewed were from issue Number 1 of 1968 to Number 12 of 1971. The major subject field searched was Behavioral and Social Sciences; the specific subfields examined were Human Factors Engineering, Man-Machine Relations, Personnel Selection, Training and Evaluation, and Psychology (Individual and Group Behavior).

In addition to these systematic searches of source listings, the act of reading in the literature unearthed other literature of relevance. Particularly valuable in suggesting articles and books of importance were issues of the *Psychological Bulletin* and appropriate chapters of the *Annual Review of Psychology*. Thus, as a result of the systematic examination of three listing sources, the utilization of other review and discussion articles which integrated much of the thinking in the subject fields, and the normal reading of the published materials of these fields, a degree of confidence can be manifested in the comprehensiveness of the coverage of this review.

Training Evaluation and Student Achievement Measurement

Training evaluation and student achievement measurement in some ways involve similar constructs, and in some ways they involve different constructs. Moreover, several different meanings have been attached to the term "training evaluation."

There are at least three major and quite different reasons for measuring student achievement. The most time-honored of these is for determining whether the student has mastered the prescribed subject matter and, hence, can be promoted, graduated, certified, licensed, or in some other way acknowledged. This type of student measurement takes place for purposes of evaluating the student; and it is completely distinct from evaluating the training provided to the student, or from other reasons for student measurement.

A second reason for student measurement is to determine his subject matter areas of strength and weakness for reinforcement and feedback purposes and for diagnosis and subsequent remedial action. Many automated, or programmed, instructional texts and devices provide for this type of measurement, as do most good tutors. This student measurement is an instructional technique, and it is completely distinct from evaluating either the student or the training.

Finally, student measurement is employed for purposes of drawing inferences about the effectiveness of the instruction provided to the student. Other things being equal, it can be inferred that the more the students have achieved, the better the quality of the instruction. Student achievement in this case is, indeed, a method of training evaluation. In only one, then, of the three uses of student measurement does student measurement overlap the topic of training evaluation. In the other two uses, student measurement is a distinct topic of interest without any necessary reference to training evaluation.

The term training evaluation also has multiple meanings and has been applied in a number of different contexts. At a minimum, one should distinguish comparative or relative training evaluations from more absolute evaluations of training. The first case involves the determination of which is best among a number of methods or programs for presenting the training content. The second case involves determination of how good the training is.

In addition to the obvious syllogistic point that a particular program may be the best and yet not be very good, the relative or absolute distinction has other implications for this review. The time frame covered has seen exceedingly rapid acceleration in the rate of development of new instructional methods. From Pressey and Skinner's early teaching machines, to a number of different approaches to programmed texts, to computer assisted instruction, the "traditional" classroom has probably undergone more of a metamorphosis in this relatively brief time period than in all of its preceding years. And, with each new development, a multitude of evaluations comparing it either to traditional methods or to the last new development have appeared. The result has been a literature very full of comparative training evaluations. No attempt has been made to discuss more than a sample of these comparative evaluations. To do more would overbalance the review with, in many cases, rather trivial studies.

The major thrust of this review is on systems, quantitative methods, and evaluations of training which have utilized more absolute criteria. Such studies have maximum import for the quality control stage within an instructional system. This quality control stage in an Air Force context is concerned with how well students are prepared for job performance, not whether the Air Force's method is better or worse than someone else's.

II. DIMENSIONS OF EVALUATION

Roles, Uses, and Characteristics of Evaluation

Stake (1969) and his associates (Stake & Denny, 1969) differentiate between evaluation and scientific research, while admitting that both can overlap. Stake indicates that evaluation studies are concerned with worth or value while research studies are rarely concerned with these issues. Stake also defines what is meant by "high" and "low" forms of evaluation. In high forms of evaluation, the results are generalizable across schools, situations, and students. In the low form of evaluation, the findings are restricted to the specific research situation because the experimental conditions are not samples of the universe of conditions. This delineation of the high and low forms of evaluation is analogous to the random and fixed-effects models referred to in statistical (analysis of variance) contexts. Nonetheless, many persons engaged in student measurement and training evaluation research have used fixed-effects designs and then erroneously generalized to other programs of instruction.

Flanagan (1969) and Bloom (1969) define what is meant by the terms "formative" and "summative" evaluation. Formative evaluation is a process concerned with the development of an educational program. Summative evaluation, though, is primarily concerned with evaluation at the end of a program. Stake (1969) feels that this distinction between summative and formative evaluation is trivial since formative evaluation never ends for the instructors and program developers. A program is summative only for someone who is outside the program and looking in for a statement of its effects.

Thelen (1969) feels that the role of evaluative measurement is ". . . feedback, diagnosis, and steering. . ." of the student. Merwin (1969), taking a broader view, thinks that there are three roles for evaluation: (*a*) school planning and administration which includes pupil classification, diagnosis of learning disabilities, appraisal of pupil progress, identification of special aptitudes, pupil promotion, and effectiveness of teaching; (*b*) instruction, its diagnosis and effectiveness; and (*c*) student decision making or helping the students to plan and evaluate their own educational experiences. Similarly, Cronbach (1963) lists course improvement, decisions about individuals, and administrative regulation as the purposes of evaluation.

Wittrock (1970) defines evaluation as making decisions and judgments about instruction and the causes of learning. It is noted that such judgments of causal relations are difficult, inasmuch as differential psychology has studied individual differences to the exclusion of cause and effect relations among learners, educational environments, and learning. The evaluation of instruction, according to Wittrock, should include observation of the student's environment (*e.g.*, teacher characteristics, student background), evaluation of the learners via achievement testing, and evaluation of learning or of permanent behavior changes. Denova (1968), using a similar paradigm, says that evaluation has three components: assessing changes in employee (student) behavior; observing whether training helps achieve organizational goals; and evaluating the training programs, techniques, and personnel.

G. Johnson (1970) lists three characteristics of evaluation: establishing merit, applications, and multidimensionality. Johnson's dimensions of evaluation are objectives, processes, components, end-products, environmental context, secondary or unplanned effects, and costs.

Angell, Shearer, and Berliner (1964) list four uses for evaluation data: (*a*) early detection and correction of behavior; (*b*) continual modification of instructional procedures when appropriate; (*c*) knowledge of whether desired achievement levels have been attained; and (*d*) acquisition of learning curves.

According to Gagne (1970), evaluation has two meanings. The first meaning of evaluation involves the determination of the worth of a system or program, and the second meaning involves determining if learning has occurred. These uses appear to be directly analogous to the topic of this literature review. Provus (1969), emphasizing training functions, thinks that the purpose of evaluation is to determine whether to improve, keep, or end a program. Evaluation is agreement with program standards, determining if a discrepancy exists in some aspect of the program, and using this information to delineate the weak points of the system.

Wiley (1970) compares and contrasts the concept of evaluation with the concepts of appraisal and assessment. According to Wiley, assessment and appraisal involve the process of ". . . judging what is valuable and ascertaining the particular levels of valued traits (p. 260)." Evaluation, though, is concerned only with the latter, and it must be empirical and behavioral. Appraisal, therefore, involves a designative and an evaluative function. Continuing, Wiley says that ". . . evaluation consists of the collection and use of information concerning changes in pupil behavior in order to make decisions about an educational program (p. 261)."

Jaeger (1970) feels that evaluative techniques can be applied to institutional decision making and educational management. Evaluation can be helpful in allocation of resources in terms of educational need, in modification of school programs, and in promotion of public understanding of the meaning of test scores.

Crawford (1969) and Berdie (1969) both have rather contrasting views of evaluation usage. Crawford feels that the goals of evaluation are increased efficiency, decreased time, and decreased costs. Berdie, though, feels that the uses of evaluation are educational, vocational, and individual.

Perhaps the best statement of the use of evaluation is given by Hemphill (1969). He says that the worth of an evaluation study is based ". . .on its contribution to a rational decision process in which it is necessary to estimate the probability of a desirable but uncertain outcome of an action

chosen from a number of alternative actions (p. 219)." In this sense, evaluation is an aid to the decision making processes.

Thus, educational evaluation has meant a number of different things to different people. The literature indicates it to be multidimensional in purposes, and these purposes seem to vary across the goals of the evaluators. Few have separated measurement (the act of deriving data) from evaluation (the judgments) made on the basis of the data. Such a taxonomy might represent at least an initial step toward providing a unifying conceptual scheme. In this sense, educational evaluation is a process which is used to make decisions with regard to instructional programs, instructors, students, institutional planning administration, and costs. Measurement represents a set of techniques which are applied to derive the data on which the evaluation is based.

Specification of Objectives

Many writers (*e.g.*, Bloom, 1969; Flanagan, 1969; Glaser, 1967, 1970; Glaser & Glanzer, 1958; Lavinsky, 1969; Peck & Dingham, 1968; Waina, 1969; Whitmore 1970a, 1970b, 1970c, 1970d) have stressed the need for a carefully specified set of objectives as a precursor to training and evaluation. While this seems self-evident, early specification of objectives often seems to be ignored. Most of the sources indicate that objectives should be defined in terms of skills and behaviors. An essential step, then, prior to the specification of objectives is a behavioral job analysis from which the basic job requirements can be derived. This process should result in a training program composed of small, discrete units with each unit having its own objective. Wittrock (1970) and Cronbach (1963) add that the specification of behavioral objectives allows absolute rather than relative student measurement. This enables one to determine who has and who has not achieved the objectives rather than who scores best or worst.

Bloom (1969) suggests that there should be consideration of the intangible outcomes of instruction. The intangible outcomes may be desirable (*e.g.*, stimulation of extra reading) or undesirable (*e.g.*, dislike of subject matter), which can lead to a revision or change in the educational objectives. These outcomes, however, seem quite amorphous and subject to considerable measurement error.

At a still higher level of abstraction, Carpenter and Rapp (1969) would determine the objectives of training by removing any objective which is dependent upon another (a concept which is theoretically neat but impractical); eliminating any objective that will not be affected by the choice of alternatives (a rather nonempirically defined concept); and finding an abstract objective to which all of the alternative objectives are means (which leaves the weighting of the alternative objectives open).

Thus, the determination and specification of objectives can assume a number of levels. These range from "objectively" derived statements of required skills and knowledges through motivational constructs and finally through complete abstraction.

Systematic Approaches to Course Development

Approaches to course development have also ranged from broad based molar systems through more discrete and molecular methods.

Carss (1969) advocates the use of a flow chart model of the educational system components in order to derive a course. This model should contain the flow of behaviors or acts needed to complete training. In the operation of the educational system, the relevant variables are identified and quantified and converted into formulae to determine the effect of output (*e.g.*, student behavior) when different inputs are considered. This is a simulation technique because one does not need to intervene in the school. In addition, Carpenter and Rapp (1969) add the obvious point that when different systems are being compared, all of their aspects which could affect output should be the same except for those being studied.

In an earlier paper, Glaser and Glanzer (1958) listed four requirements for course development:

1. *Specification of objectives*–A list of the objectives of the course in behavioral terms.
2. *Input control*–The selection of enrollees into the training program (*e.g.*, number of men available, testing costs, etc.)
3. *Techniques and methods of training*–Decisions regarding the amount of practice, learning guidance, reinforcement, extinction, training sequence, meaningful relationships in learning, use of punishment, learning plateaus, motivation, individual differences, etc.
4. *Output control*–Measurement of training (*e.g.*, formative evaluation, setting of proficiency standards, diagnosis of training inadequacies, performance tests, etc.).

Osborn (1970) presents an interesting model which he calls a "closed loop" approach. However, as early as 1950, workers in the area have regarded training evaluation to feed back to the instructional process. Thus, the closed loop concept would not be regarded as a "new" development. Osborn indicates that job requirements lead to training objectives which result in training content and performance tests which ultimately yield an evaluation of the quality of student performance in terms of job requirements. Osborn feels that it is often too costly to develop a full field performance test for a large number of individuals. He suggests a matrix approach as the solution to this dilemma. First, the job components (behaviors) are listed across the top of the page. Down the left side of the page is a list of the potential test methods graded in degree of complexity from full field to paper-and-pencil (*e.g.*, simulations, photos, pictures, drawings). Osborn contends that many times it is necessary to compromise—to sacrifice relevance and diagnostic capability for economy. The alternatives must be considered, and then the most complex, yet feasible method, must be selected and used.

The sequence of course development used in the Army's Trainfire I program (Crawford, 1969) includes (*a*) job analysis; (*b*) transfer of the job description into a test of how well the man performs the necessary skills; (*c*) development of new training stressing realism, clarity, and simplicity; and (*d*) experimentation using a conventionally trained group and an experimentally trained group which are compared on the test.

Glaser (Glaser, 1970a, 1970b; Glaser & Cox, 1968) presents a somewhat more elaborate model than his earlier version (Glaser & Glanzer, 1958). This new model includes the following:

1. Specification of objectives in terms of observable behavior. Criterion-referenced measures indicate the content of the subject's behavior in regard to the objectives and without regard to the performance of others.
2. Diagnosis and profiling of the subject entering instruction. The types of entering behavior that need measurement are previous extent of achievement in the subject area, prerequisites, learning set variables, ability to make discriminations, and general intelligence.
3. Selection of "instructional alternatives" based on the diagnositic and profiling step of the system.
4. Continuous assessment and monitoring which can include frequency of correct answers, errors in relation to a standard, speed, transfer and generalization, attention span, and response latency.
5. Adaption and optimization. The treatments and individual differences may interact; therefore, individuals should be adapted to the best treatment. Those that interact most with the treatment are the most important. Decisions about treatments should be made sequentially, and these should be optimized by using quantitiative methods.
6. Evolution or self-contained improvement capability that modifies itself after acquisition of new knowledge.

A system which mirrors much of the prior thinking is the Instructional System Development (ISD) technique developed by the United States Air Force (Air Force Manual 50-2, 1970). This system in its latest form contains the following steps:

1. Analyze system requirements
2. Define education or training requirements
3. Develop objectives and tests
4. Plan, develop, and validate instruction
5. Conduct and evaluate instruction

Hunter, Lyons, MacCaslin, Smith, and Wagner (1969) feel that training program content must be job relevant. Taking the seven-step Human Resources Research Organization method of curriculum development and applying it to what the services are doing, they reported several findings: (*a*) System analysis for training purposes was not used in any of the services; (*b*) there was a requirement for task inventories in the Army and Air Force; (*c*) there was no development of a job model for any service; (*d*) there was no task analysis for curriculum development; (*e*) all services said training objectives should be job relevant but no provision was made for specificity; (*f*) training program development procedures were not maximally effective because the objectives were not fully specified; (*g*) very little or no evaluation and assessment of training effects (the Air Force had the only standards of graduate behavior and was the only service to perform field visits); and (*h*) training accounted for 6 percent of the defense budget.

In summation, the systematic approaches to course development attempt to account for almost

all of the variables that can affect training and student behavior. Most of the systems begin with job analysis in order to derive a set of behavioral job requirements from which training objectives can be formulated. Many writers advocate a pre-training assessment of the entering students in order to channel them to the training program which is most suited to their needs and abilities. Performance tests and other measures of student behavior are then constructed in order to reflect the training objectives. Finally, after training the students, the training programs are evaluated through various means.

Measures and Methods of Evaluation

Campbell (1971) presents a rather dim picture of the current state of methodology in training and evaluation literature. He feels ". . . by and large, the training and development literature is voluminous, nonempirical, nontheoretical, poorly written, and dull (p. 565)." Continuing, Campbell says that ". . . In sum, the methodology of training and development research cries for innovation. . . . As yet we have no workable technology that is capable of producing a large amount of training research data (p. 579)."

Similarly, Schultz and Siegel (1961a, 1961b) as the result of a comprehensive review, observed earlier a need for a unifying conceptual structure with more emphasis on theoretical development in the area of job performance rather than technical advancements. They argued for more research based on an integrative theoretical framework rather than on an inductive framework.

Campbell, Dunnette, Lawler, and Weick (1970) divide training criteria into two groups. Internal criteria are those directly concerned with the training itself, while external criteria measure post-training or on-the-job behavior. These authors recommend the use of multiple criteria, each reflecting different aspects of the organization's goals. Gagne (1970) presents a similar dichotomy in which he stresses initial problems directly connected with the lesson and transfer problems involving principles taught in the lesson.

Use of a composite overall criterion will undoubtedly obfuscate important relationships since many of the subcriteria within the composite are probably orthogonal (Cronbach, 1963). According to Dunnette (1963), it is preferable to have multiple criteria in order to account for a greater proportion of the behavior variance.

The evaluation or measurement must not be affected by the method of measurement or research procedure. Even the presence of the experimenter or the process of evaluation itself can alter the results (Bloom, 1969; Cronbach, 1963). According to Gagne (1970) two evaluation criteria for measures are "distinctiveness" and "freedom from distortion."

Weiss and Rein (1970) claim that broad based evaluation programs have design and technical problems so ponderous as to make any evaluation impractical and questionable. They propose a developmentally oriented, more qualitative evaluation as being more appropriate. Weiss and Rein imply that where there are many variables to consider, one can not possibly prove or disprove the values of any program.

Biel (1962) says that ". . . fundamental criteria for evaluating a simulation-based training program or device is the extent of transfer of training to the live situation. . . . In cases . where ultimate criteria are obviously unavailable, intermediate criteria must be employed. One example of an intermediate criterion is performance in a final examination. . . Sometimes improvement as measured by performance on the training device itself is the best measure available of the effectiveness of the device and its associated training program (pp. 377-378)." Gagne (1968) has given a similar emphasis to transfer of training.

Crawford (1962) and Glaser and Klaus (1962) posit that proficiency tests developed from job analysis should be employed to evaluate students and training. The standards on the proficiency test must be based on acceptable or adequate job behavior.

Cronbach (1963) feels that, in training evaluation and student measurement, the testing of terminology which is specific to the training course should be kept independent from tests of understanding of content. A person who is not taking the course should be able to understand (not necessarily answer) the question. Cronbach also classifies transfer of learning into an immediate and a long-term category. Immediate transfer involves testing the student's course knowledge, while long-term transfer is concerned with aptitude gain and learning to learn.

Angell, Shearer, and Berliner (1964) list three types of training measures:

1. *Initial measures* given prior to instruction or training and which are used for selection purposes. The correlation between the selection tests and future performance should be high.

2. *Interim measures* taken while training is in progress, and ". . . they are more accurately predictive of terminal proficiency than are measures made earlier (p. 3)."

3. *Terminal measures* obtained after training is completed and which are predicted by the initial and interim measures. Some examples of terminal measures are written tests, oral tests, performance tests, expert judgments, and rating scales.

Peck and Dingman (1968) present a unique method of evaluating student teachers. Training is attained when each of the training objectives is reached by the student teacher, and these advances yield significant pupil gains in the classroom.

Della-Piana and Berger (1970) have provided a design for conducting pilot studies on the efficiency of programmed instruction. They begin with six to eight subjects of above average ability who can give verbal feedback which is relevant to program revision. The subjects are split into groups of three or four each. The groups are presented with the programmed instruction, and, on completion of the training, they are queried regarding possible revisions for the program.

Thelen (1969) describes diagnosis (progression toward goals) and troubleshooting (difference between what exists and what ought to be) in the context of group instruction. In group instruction, the students are unsupervised most of the class time, and the instructor can only hope to sample their behavior. In a highly structured class, the evaluation is in an authoritarian framework in which student and teacher behavior are evaluated on several continua from good to poor. This can be considered evaluation of deviancy. In the unstructured class, no set of criteria for describing deviant behavior can exist. All behavior is thought to be relevant, and attempts are made to account for it, or to understand why it occurred. The authoritarian teacher knows what is to be taught and determines the extent to which individuals differ in meeting expectations. The more democratic instructor will use games, ungraded classes, small work groups, and student cohesiveness. Finally, Thelen advocates the use of "barometric" individuals, or students who respond consistently and selectively to instruction or to some other important group condition.

Wiley (1970) advocates a system of evaluation which could lead to a great savings in time. First, if all the students in the class receive the same experimental treatment, then the appropriate statistical datum is the class, not the student. When the datum is a collective, one can sample from it and save considerable time. In addition, one does not have to give each student all the items. Even single items can be used, and they are easier to interpret than total scores. Jaeger (1970) uses the aforementioned sampling strategy for institututional decision making.

Wiley also introduces some new terminology in his descriptive system of evaluation. First, the *standards* of evaluation involve designating traits to evaluate and designating the levels that are thought to be appropriate. Secondly, the *object* of evaluation is the instructional program and its component parts. Next, the *vehicles* of evaluation are directly affected by the objects, and they consist of students, classes, or schools. Finally, the *instruments* of evaluation display the behavior of the vehicles. Wiley says that the fundamental problem in evaluation ". . . . is to establish the effects of the objects on the vehicles by means of the instruments (p. 262)."

Furno (1966) has an evaluation approach confined to educational surveys. The sequential elements in Furno's system are (*a*) specification of survey objectives; (*b*) definition of the population; (*c*) description of what information is to be collected; (*d*) determination of the best mode of measurement; (*e*) selection of the sampling unit; (*f*) selection of the sample; (*g*) planning of field work so that it will be carried out smoothly; (*h*) conduction of pilot study; (*i*) provision for data processing; (*j*) analysis of data; and (*k*) storing of survey information and providing for access when needed.

Somewhat less elaborate are, Hawkridge's (1970) seven phases of evaluation research: (*a*) specification of objectives; (*b*) selection of objectives to be measured; (*c*) selection of instruments and methods; (*d*) sample selection; (*e*) measurement and observation schedule development; (*f*) choosing analytic techniques; and (*g*) drawing conclusions and making recommendations.

Campbell (1970) suggests a completely selective approach including the use of an evaluation model which measures trainee reactions, trainee learning, trainee behavior on the job, and results with regard to the organization. Campbell concludes that too many evaluation studies have focused on the measurement of trainee reaction (*e.g.*, attitudes and opinions), to the exclusion of the other dependent measures.

Flanagan's (1969) system of evaluation includes (*a*) defining the outputs of the system including

the objectives and unplanned effects; (*b*) selecting the procedures needed to measure the worth of the outcomes (*e.g.*, costs, benefits); and (*c*) composing a plan based on analysis including a decision and overall evaluation of the final program.

Possibly, an evaluation which aims to be at all complete should include consideration of most, if not all, of Scriven's (1967) criteria. They include (*a*) knowledge of specific items of information and patterns and sequences of information items; (*b*) comprehension of internal relationships within the field (*e.g.*, inferences and implications), interfield relationships or the association between the knowledge of one field and that of another, and application of the field or its principles to an appropriate example; and (*c*) motivation and attitude toward the course, the subject, the field, field relevant materials, learning and knowledge activities in general, school, career teaching, the teacher, peers, and self.

Problems of Evaluation

As was mentioned previously, Campbell (1970) thinks that too many evaluation studies use measurement of trainee reactions to the exclusion of trainee learning, trainee behavior on the job, and effects on the organization. Trow (1970) feels that much innovation in training is done for its own sake to relieve boredom and only secondarily for its outcomes. Evaluation studies are too often large-scale and aimed at funding agencies to prove that the innovation is of value.

C. Harris (1970) points out that most investigators fail to integrate prior research into their experimental designs. He goes one step further by posing the question of integrating prior-research findings into numerical research analysis. Harris' concept would be feasible if more collaboration could be achieved among different agencies and investigators. A related problem (Lortie, 1970) is whether or not ultimately too much centralized evaluation will be achieved (without realizing it) through the use of computers and data processing equipment. Clearly, an optimum middle ground must be found.

Student measurement can have both positive and negative effects. The person being evaluated will always respond to evaluation in terms of the perceived fairness. If he perceives the evaluation as unfair, the person being evaluated may become resentful, especially if the evaluation is more critical to his career or to his student status (Bloom, 1970).

Evaluation cannot function in an authoritarian society which resists social change. Evaluation also does not function well in an equalitarian society because all persons in it are considered equal. In actuality, evaluation functions best in a competitive society (Berdie, 1969). One must also consider the various publics at which the evaluation is aimed. These publics are trainees, trainers, sponsoring organizations, training technicians, and social scientists. The value of a particular type of training must be presented to the public with which it is concerned, and it may be different for each public (Bass, Thiagarajan, & Ryterban, 1968).

Walker (1965) performed a study illustrating one of the most serious problems in evaluation research. He asked 20 training experts to rate 16 training techniques with regard to 34 training selection criteria. These training personnel tended to select training methods based on administrative and contractual needs to the exclusion of training methods based on educational and psychological principles. Walker concluded that this group of training experts was more concerned with budget and training time than with learning.

Berdie (1969) lists conceptual needs and problems of evaluation and measurement. He identifies the requirement to evaluate whole persons and the various ways in which traits cluster together; and, further, the need to know more about statistical as opposed to clinical prediction. Breadth of evaluation in addition to depth of evaluation must be considered; and various statistical modes of prediction must be attempted (*e.g.*, moderator variables).

Smode, Hall, and Meyer (1966) severely criticize Air Force evaluation research. They contend that (*a*) different dependent measures are often used across studies leading to incomparability of results; (*b*) too much stress is placed upon subjective opinions (*e.g.*, rating); (*c*) different limits or standards are used for describing performance; (*d*) too many personnel and equipment changes occur during the execution of many studies resulting in a lack of proper research control; (*e*) different methods of processing and interpreting the transfer of training data are employed; (*f*) presentation of the same study in different reports makes it difficult to determine exactly what was done; (*g*) inadequate and imprecise criteria are used; (*h*) comparability and control of skill levels of subjects and trainees are lacking; (*i*) there is difficulty in matching research criteria and tasks to flight conditions and demands; and (*j*) there is disorganization and lack of cooperation among researchers.

In a somewhat different context, Suchman (1967) presents a systematic overview of the shortcomings of evaluation research in general. First, with regard to objectives, Suchman feels that certain excesses have tended to characterize the research: too much arbitrary problem selection; too much stress on resources and material and not enough on achievement; too much stress on quantity of services and record keeping at the expense of true evaluation; too much emphasis on program objectives based upon tradition and common sense; too much mixing of final, intermediate, and immediate objectives; and too much idealism and not enough realism.

In listing inadequacies regarding procedural methods, Suchman criticizes the excessive emphasis on research based on available or existing records which discourages the gathering of new data; the absence of sound experimental designs, thus making it difficult to determine if change is the result of innovation or chance; the use of measurements of unknown consistency and accuracy; the use of weighting methods and standards too often based upon rational rather than empirical means; the inadequate allowance for or control of demographic variables (*e.g.*, locale, race, age) making interpretation difficult; and the over-emphasis on correlation with inadequate attention to causality.

Suchman also comments on the administration of evaluation studies, contending that evaluation guides are too often used by unsophisticated persons, thus making analysis and comparison of ratings difficult. Further, he suggests that self-evaluations are too often used, which allows bias to contaminate data. And, finally, when supervisors are forced to perform evaluations in addition to their usual activities, it becomes difficult to properly plan, organize, and conduct evaluation studies.

What generalization can be extracted from this mass of critical rhetoric? First, these writers seem to think that there has been too much use of rational (armchair) rather than empirical methods. Similarly, they feel that evaluation research is too often subjective when objectivity is needed. Finally, evaluation research is too often limited by monetary considerations. The monetary criticism is probably the most important, since most of the other criticisms can be reduced to it. What most investigators do not realize is that cost cutting actually wastes money because the results of the research are at best uninterpretable. Many agencies, contractors, and others doing research might be well advised to save their money and do perhaps one or two sound research studies rather than five or six poor ones.

Summary

In the first section of this chapter, the roles, uses, and characteristics of evaluation were discussed. Evaluation was differentiated from research. Formative and summative types of evaluation were discussed. Also, evaluation was contrasted with appraisal and assessment. It was concluded that evaluation is a process which is used to make decisions with regard to instructional programs, instructors, students, institutional planning, administration, and costs.

The second part of this chapter contained a short discussion of objectives. Most of the sources reviewed seemed to indicate that each unit of training must have a behavioral objective based on the job requirements.

The third portion of this chapter contained a systematic overview of approaches to evaluation and course development. These systems approaches to evaluation and course development attempt to account for almost all of the variables that can affect training and student behavior.

The fourth segment of the chapter consisted of a discussion of the measurement aspects of evaluation. There was a presentation of the various types of criteria that can be used in evaluation studies. Emphasis was placed on the multidimensional aspects of criterion measurement. Most of the writers reviewed suggested that transfer of learning was the ultimate goal of training. Also, sampling procedures were suggested as a means of saving time and costs when the units of measurement are whole classes and schools.

The final section of this chapter presented a discussion of the various problems and difficulties involved in evaluation studies. Several conclusions were drawn:

1. There is too much use of rational rather than empirical methods.
2. There is too much subjectivity when objectivity is needed.
3. Evaluation research is too often limited by monetary considerations.

III. QUANTITATIVE METHODS AND DEPENDENT MEASURES

Characteristics of Dependent Variables

Fitzpatrick (1970) lists four characteristics of criteria which he thinks are essential for any evaluative measure. First, the criteria must be relevant to the objectives being measured. Second, the criteria must be comprehensive and cover all

important objectives. Third, the criteria must be reliable within the limits of cost. Finally, the criteria selected must be feasible, and this is determined almost solely by cost.

Bloom (1970) also makes a set of very relevant comments concerning validity with regard to student measurement and training evaluation. Generally, content validity is stressed in training evaluation, while construct validity is emphasized in assessment and appraisal. Student measurement, though, usually emphasizes predictive and concurrent validity. Bloom feels that the type necessary should be determined and not be confined to one or another. Bond and Rigney (1970) add that the dependent measure which "best predicts final performance" should always be selected.

Several indices may be related to final performance, and the computer can be used to choose and weight them.

Gideonse (1968) lists several types of measures that can be used for measuring students and for training evaluation. Gideonse's measures are (*a*) student achievement as measured by tests (which leaves many of the student's intellectual qualities untapped); (*b*) a desirable change after a stimulus input; (*c*) dropout or attrition rate; (*d*) attitudinal and motivational measures; (*e*) education levels; and (*f*) facilities, equipment, materials, human resources, pupil expenditure, non-school activities, organization patterns, and administrative agencies.

Campbell and Dunnette (1968) add that most T-group research involves the use of attitude scales or opinion change as criteria rather than organizational performance or improvement.

Crawford (1967) indicates that proficiency tests, when used to evaluate training programs, should not just be used at the end of training, but should also be used to test retention after a period of disuse. Similarly, Martin (1957) divides criteria into those based on the content of the training program (internal criteria) and those based upon job behavior (external criteria).

Englemann (1968) contends that there are two kinds of conditions which can indicate that learning has occurred. In the *fixed* condition, a response or instance of behavior is used to show that learning has taken place. This is the criterion of performance. In the *variable* condition, several responses can show that learning has occurred. One can easily see that within this latter condition, it is easier for the student to demonstrate that he understands the concept being taught since the requirement for learning in the variable condition is dependent on a concept or rule and not on a response. Englemann adds that both the fixed and variable conditions are needed depending upon the situation.

Kelley and Kelley (1970) document a unique type of dependent measure for research which holds the traditional dependent variables of speed and accuracy constant. They work with an "adaptive variable" which is the adjustment the student must make to obtain a certain score with speed and accuracy held constant. The adjustment is the dependent variable, and it can be any variable which affects performance.

Test Construction

Denova (1968) lists the steps in test construction as follows: (*a*) defining test scope, (*b*) defining what is measured, (*c*) choosing items, (*d*) choosing the most appropriate testing technique, (*e*) determining the number of items, (*f*) choosing final items, (*g*) arranging items, (*h*) writing clearly understandable directions, (*i*) constructing a scoring template, and (*j*) evaluating questions. Evaluation of the test, of course, involves such factors as (*a*) validity, (*b*) reliability, (*c*) simplicity, (*d*) distribution, (*e*) content, (*f*) objectivity, and (*g*) difficulty level. Other, more exhaustive, accounts of test construction and its concomitant problems can be found in many sources such as Air Force Manual 50-9 (1967), Gronlund (1968), and Wood (1960). The remaining parts of this chapter, therefore, are devoted to some new techniques and applications.

Horn (1966) feels that a predictor test must have internal consistency in order for it to correlate adequately with a criterion. On the other hand, he feels that assessment tests need representativeness of content regardless of internal consistency. He demonstrated that his own classroom assessment devices were more like predictors than assessors. Horn concludes that there is no reason why assessment devices must have low internal consistency reliability.

McGuire and Babbott (1967) constructed a test for medical students consisting of a series of simulation exercises. The test begins with a case write-up and several possible courses of action or diagnoses. Each choice the student makes is branched to other choice points until the patient is either dead, transferred, or gets well. In the construction of the test, a panel of experts rated each choice along a five-point scale which ranged from "clearly indicated' to "clearly contra-indicated." Several possible scores result from the procedure. The *efficiency* score is the percentage of the student's answers which are helpful to the patient.

The *proficiency* score is the percentage agreement with the criterion group (optimal patient care). Proficiency, then, is a combination of errors of commission and errors of omission. The *composite* score is a function of proficiency and efficiency. According to McGuire and Babbott, traditional multiple-choice tests take a portion of behavior and treat it independently of the total behavior pattern of which it is a part. This stresses "product" as opposed to "process." McGuire and Babbott conclude that their test stresses the process aspects of behavior and that it is uncorrelated with most multiple-choice tests.

Westbrook and Jones (1968) used a class of psychology graduate students to construct a multiple-choice test of Anastasi's testing book. There were 54 items in form A and 54 items in form B. The Kuder-Richardson reliability was .73 and the split-half reliability was .62. The tests were validated against a teacher-made test, resulting in validities of .75 for form A and .59 for form B. Evidently, graduate students can be used to construct fairly reliable and valid tests.

Gorth and Grayson (1969) developed a Fortran computer program which can "...compose and print any number of tests consisting of questions, multiple-choice or completion type, selected from an item pool (p. 173)." This program will make as many copies as is desired, randomize multiple-choice answers, and print scoring keys. Apparently, this program is for sale.

Forrest (1970) wished to develop an objective flight test for private pilot certification. His test consists of a miniature sample of flying situations typically met by pilots. Each situation involves an evaluation and an action. The test measures (*a*) retention and recall, (*b*) judgment, (*c*) planning and problem solving, (*d*) perceptual-motor coordination, and (*e*) habit. The actual test was a cross-country flight with a pre-flight and an in-flight phase ($N = 15$). Scores on the test correlated .50 with expert ratings.

Hierarchical and Sequential Testing

Hierarchical and sequential tests involve a sequence of branching in which the student only gets items at his own level. This procedure decreases testing time, increases reliability, and increases student motivation because he is not forced to take and guess at the more difficult items. The concept was introduced in early "intelligence tests" and has recently received new emphasis. An example of the application is the work of Cleary, Linn, and Rock (1968a, 1968b) who wished to use programmed tests to decrease testing time while leaving reliability and validity the same. In the procedure described by Cleary, Linn, and Rock, each student receives a different set of items along a scale. Sequentially programmed tests have a routing section which branches the subject to the appropriate items and a measurement section containing items of suitable difficulty. The routing section can be used alone, although these investigators used a combination. These authors used the test scores of 4,885 11th grade students on the School and College Ability Tests (SCAT) and Sequential Tests of Educational Progress (STEP). The sample was divided in half, with the second half used for cross validation purposes. The subjects in the initial validation effort were routed into four groups using four different sequential sampling procedures. One of the four routing methods, the sequential method, produced the fewest errors of classification and the highest overall correlation with the total SCAT and STEP test scores. The sequential method uses fewer items for those easy to classify and more items for those at the borderline of categories. The measurement test is constructed by obtaining the items with the 20 highest within-group point-biserial correlations (excluding the routing items). Computer based testing could facilitate this procedure because of speed, flexibility, convenience, and immediacy of feedback. This method is especially suited to persons at the extremes of the distribution because they can be quickly routed and thus save time. One problem acknowledged by the authors, with this research effort, is that the SCAT and STEP items were taken out of context from a total test. This could have biased the results.

Lord (1971a, 1971b) introduces a theoretical treatment of "tailored testing" which is a sequential testing procedure consisting of one rather than two stages. It is tailored in the sense that the items are those that are best suited to the individual being tested. "In tailored testing we try to choose items for administration that are at a difficulty level that matches the examinee's ability, which we infer from his responses to the items already administered. . . . when the examinee gives a wrong answer to an item, the next item administered should be an easier one; when he gives a correct answer the next item administered should be harder (Lord, 1971a, pp.3-4)." In his earlier work, Lord (1969) evolved a two-stage testing procedure using similar principles.

Ferguson (1969) used a computer to select items on the basis of a student's prior responses. The computer will keep testing the student until

he satisfies the criterion specified by the training objective. When the criterion is met, the computer will route the subject to the next training objective containing items based upon the student's proficiency on the first training objective. The program was successfully used with 75 elementary school students from the Pittsburgh area.

According to Gagne (1967), if the curriculum units are arranged hierarchically, and the test items meet standard requirements, a hierarchical testing procedure will be implicit since most people who fail the lower unit will not pass the next higher unit. Moreover, if persons who pass a lower unit fail on the next higher unit, an additional interspersed unit may be indicated. Obviously, this technique can also indicate whether or not some units have been reversed in the hierarchy of instruction.

Criterion- and Norm-Referenced Testing

Glaser (1963) and his colleagues (Glaser & Cox, 1968; Glaser & Klaus, 1962; Glaser & Nitko, 1971), as well as Popham (1969), Carver (1970), and Holtzman (1971), have all written on the topic of criterion-referenced versus norm-referenced testing. The characteristics of criterion-referenced tests are that they (*a*) indicate the degree of competence attained by an individual independent of the performance of others; (*b*) measure student performance with regard to specified absolute standards of performance; (*c*) minimize individual differences; and (*d*) consider variability irrelevant.

Generally, from these statements, it can be seen that criterion-referenced tests tell how the student is performing with regard to a specified standard of behavior. Individual differences are considered irrelevant, since the student is graded against a single standard rather than against all the others taking the test. Assigning grades of competence to students on the basis of relative performance, when it is not really known whether any of the students have attained a specified behavioral objective, makes very little sense. One can, though, derive individual differences from criterion-referenced tests by specifying the degree of competence reached by each student.

Simon (1969) thinks that there is no real difference between criterion- and norm-referenced tests. Whether a test is one or the other depends upon how the scores are used.

Glaser (1963) and Glaser and Cox (1968) discuss the use of norm-referenced achievement tests and criterion-referenced tests in differentiating among individuals and treatment groups. When evaluating individuals, one needs to use an achievement test containing items with different difficulty levels. For evaluating treatments or experimental conditions, though, one needs perfect post-treatment answers and incorrect pre-treatment answers so that the dependent measure is maximally sensitive to training change. In this latter case, criterion-referenced tests are most appropriate.

K. Johnson (1969a, 1969b) suggests that training evaluation should use criterion-referenced tests, but that they are costly and just not feasible for many training situations. Johnson's purpose was to determine the degree which other measures (*e.g.*, norm-referenced tests, student and instructor attitudes) can be used as substitutes for criterion-referenced tests. Reliabilities were calculated for three measures on four courses taught at the Naval Air Technical Training Center. In one course there was a comparison with criterion-referenced tests. The reliabilities for all three methods were fairly high, but a large number of items was needed (*i.e.*, more than 20) to get an adequate reliability for norm-referenced tests. Student and instructor attitudes were highly correlated, but neither had a high correlation with norm-referenced tests. Each of the three measures accounted for 27 to 43 percent of the variance of scores on criterion-referenced tests. Without defining what he considered to be an adequate substitute, Johnson concluded that none of the other methods is an adequate substitute for criterion-referenced tests.

Siegel, Schultz, and Lanterman (1964) and Siegel and Fischl (1965) sought to develop a criterion-referenced evaluation scheme for the Navy electronics technician rating. What is unique and interesting about these studies is that the criterion referencing was done in combination with Guttman scaling procedures. Their technique involved (*a*) assembling statements of the specific system objectives of Naval air electronics; (*b*) weighting these objectives on the basis of the importance of their respective contributions to system requirements; and (*c*) psychophysically establishing cut points on a Guttman-type job performance scale, the cut points representing levels of skill required in order for each of the objectives to be met. The resultant Technical Proficiency Checkout Form Scales (TPCF) were found to correlate between .65 and .74 with performance test scores.

Ratings

Rating, although widely used, is one of the most unreliable, biased, and contaminated methods for evaluating performance. Several factors which can contribute to poor or inadequate ratings are (*a*) friendship, (*b*) quick guessing, (*c*) jumping to conclusions, (*d*) first-impression responses, (*e*) appearance, (*f*) prejudices, (*g*) halo effects, (*h*) errors of central tendency, and (*i*) leniency. Of these, the last three are probably the most important. Halo exists when a rater allows his overall, general impression of a man to influence his judgment of each separate trait on the rating scale. Errors of leniency occur when a rater tends to use only the upper portion of the rating scale when rating all or most of his men. Errors of central tendency occur when the rater uses only the middle portion of the rating scale when rating his men. Considerable evidence exists which demonstrates that rater training can reduce these sources of bias so that the resultant ratings are at least minimally useful (Bergman & Kujawski, 1969).

Howard and Correll (1966) wanted to determine if there was a consensus with regard to the acceptability of various behaviors of psychological interns among those responsible for training them. The trainers were given a list of 27 critical incident statements and were asked to indicate whether the behavior described in the incident was characteristic of a beginning trainee, an intermediate trainee, or a senior trainee. In many instances, university based trainers used more lenient standards, and in other instances agency trainers used more lenient standards. There was, of course, some agreement across universities and agencies. Overall, some behaviors thought to be characteristic of beginners in one place were thought to be characteristic of senior trainees in another place. The authors concluded that more uniformity is needed because of the widely differing standards of behavior.

In another study, Edwards (1968) had the teachers from five nursing schools rate the performance of 55 of their senior nursing school students on their performance under three conditions: (*a*) situations requiring interpersonal physical care; (*b*) situations needing technical skills; and (*c*) conditions requiring non-physical, interpersonal patient care.

Evaluations were made by the operating room instructor, the medical nursing instructor, and the psychiatric instructor. All trainees were rated from A to E. The results showed that all interrater correlations were very low (.5 at most). The only fairly high correlations were within instructors across specialties. The authors indicate that these unreliable results were caused by (*a*) teacher personality, (*b*) relations with students, (*c*) differential behavior of students, and (*d*) differential teacher criteria. The ratings also had a disappointing relationship with test scores and grades within specialty. The ratings correlated −.01 to .27 with test scores and .20 to .49 with grades.

Greer, Smith, and Hatfield (1967) constructed a standard system of checkpilot helicopter evaluation in order to overcome effects of the checkpilots' proclivity to rate on the basis of their own personal standards rather than on student flying skill. First, the training program was evaluated in terms of maneuver components. Specific proficiency scales and instrument observation were used as criteria instead of the checkpilot's own schema. From this early work the Pilot Performance Description Record (PPDR) was constructed. The PPDR consisted of items reflecting the most critical aspects of each maneuver. Fifty intermediate and 50 advanced helicopter students were each given checkrides with one research staff member and one checkpilot. Prior to this, some of the checkpilots were trained in the use of the PPDR to reduce checkpilot differences in scoring standards. The results showed that (*a*) the reliability of flight proficiency evaluations improved; (*b*) the PPDR recorded specific student deficiences; (*c*) checkpilots trained in use of the PPDR were more consistent in their evaluations than checkpilots who were only oriented in the PPDR; and (*d*) checkpilot training is necessary when using the PPDR.

In another study, Greer (1968) wished to increase the reliability of checkpilot ratings which typically averaged .20. Checkpilots were asked to complete an 11-item rating form. Those who agreed with an r of .90 or better were paired together with students; the resultant correlation was .65.

Duffy (1968), Duffy and Jolley (1968), and Duffy and Anderson (1968) wished to develop an objective recording device to score student helicopter checkrides. The students were scored during and after training and on maneuvers. All data were recorded on IBM cards, and a class percentage error and a school average were tabulated. If certain types of errors tended to show up under one instructor in one aspect of training, the instructor was given additional instructional training. If one checkpilot was found to be more strict than the other, he was also given counsel to make his ratings less strict.

Caro (1968) undertook a study to compare grades given by checkpilots and grades given by instructors before and after innovations in rating were introduced. A second study was performed to determine if grades were influenced by the checkpilot's relationship with the students or the instructors. To eliminate bias due to prior knowledge, 40 of 60 subjects were given checkrides by checkpilots outside the classes studied. The principal results of concern from these two studies suggested that (*a*) there were high correlations between instructors and checkpilots from the same classes; (*b*) there was no relationship between instructors and checkpilots from outside the classes; (*c*) student grades were affected by the individual standards of the checkpilot; (*d*) specific information was collected by the checkpilot on the student's flight, but not systematically or consistently; and (*e*) there were no differences after the new grading procedures were introduced.

Jenkins, Ewart, and Carroll (1950) sought to develop an index of combat effectiveness against which tests could be validated. They used the nomination technique which asks each man to name two with whom he would like to fly wing and two with whom he would not like to fly wing, together with the reasons for his choices (checked off on a 22-item checklist). Data were collected on 2,274 high and 1,829 low and 228 mixed pilots. The results showed that the nominations were related to the rank of the officer and that their reliability was .80. The reasons for the nominations were more reliable for the lows than for the highs. Also, there was a different frequency of use of reasons for different ranks (e.g., senior officers more often avoided going on combat missions than junior officers). A factor analysis of the checklist data delineated several underlying factors: (*a*) sociability, (*b*) practical intelligence, (*c*) coolheadedness, (*d*) combat aggressiveness, (*e*) flying skill, (*f*) teamwork, (*g*) leadership (highs only), and (*h*) reaction to failure (lows only). A second order factor analysis resulted in two high factors (fighting ability and capacity for combat leadership), and three low factors (emotional inadequacy, fear-impulsive foolish, and lack of practical intelligence). All of the aforementioned factors were orthogonal. Those interesting results notwithstanding, the ratings failed to predict combat success, even with rank controlled.

In another study, Yellen (1969) used co-worker or peer ratings as criteria of performance for field artillery crewmen. The multiple correlation between these ratings and a weighting of the major areas of a proficiency test was .71.

In one final study (Flaugher, Campbell, & Pike, 1969), white and black medical technicians were rated on job performance by both white and black supervisors. White supervisors tended to rate the whites slightly higher than the blacks, while black supervisors rated blacks considerably higher than whites.

In summation, ratings tend to improve to the extent that the influence of the rater's own idiosyncrasies are prevented from affecting his observation of subordinate behavior. The evaluator must observe and record behavior in objective terms. If this suggestion seems mechanistic and devoid of rater influence, it is meant to be that way. The more the rater can become like a behavioral metering device, the less likely he will contaminate the evaluation. Also, it will help immensely if the rating items are couched in behavioral rather than in relative or evaluative (*e.g.*, above average) terms. Finally, performance evaluations should not be tied to salary review unless they are to be used for that purpose.

In general, ratings are much used and convenient although they are at best a haphazard method of evaluating training performance, student achievement, or job behavior. If other, more objective methods are feasible, they should be used.

Cost Effectiveness

Alkin (1970) has written an extensive treatise on cost-benefit analysis. Some of his comments and suggestions are reviewed in the ensuing paragraphs.

Generally, cost-benefit analysis is the analysis of the costs and benefits of various alternative courses of action. The decision maker selects the method giving the largest yield at a given cost, or the most benefit for the least cost. Input and output must be measured in dollar terms. Cost-benefit studies are usually large-scale. For instance, the costs of college education can be compared with the resultant increase in productivity yielded by the college education.

The manipulatable characteristics are the conditions whose variations maximize or minimize student output. The manipulatable characteristics which affect student output are (*a*) student inputs measuring the achievement starting point of the student; (*b*) financial inputs or the funds allocated; (*c*) external system which is the giver of inputs and the receiver of outputs (*e.g.*, society); and (*d*) instruction, supplies, tests, and similar items.

With regard to the outcomes of cost-benefit analysis, the analyst's interest is in how the student has changed in short- and long-term ways (*e.g.*, how well he deals with other schoolwork and his society). Although there are financial inputs, there are no financial outcomes except those derived from behavior changes. There are also non-student outcomes which comprise items such as teacher salaries and number of personnel used in the program.

Alkin sees three major problems in evaluating the cost effectiveness of manipulatable variables. They include (*a*) difficulty in getting accurate cost data; (*b*) difficulty "...in dealing with cost-effectiveness estimates in the light of system-interrelationships (p. 235);" and (*c*) problems in generalizing to specific individual cases.

Hawkridge (1970) says that there are two evaluation loops regarding money allocated for educational programs. These two loops are the "philanthropic" and the "conservative." As soon as money is allocated, many programs spring up. If the evaluation is done poorly or unreliably, then the money is cut back and the first thing the program administrator usually does is cut evaluation cost so he can keep other aspects of the program. One can, of course, stay in the philanthropic loop if sound evaluation is performed.

Gubins (1970) performed a cost-benefit analysis of training programs for the hard core unemployed. In this case, cost-benefit analysis is based on the cost of unemployment and the gain from investment in these human resources. Gubins' findings suggested the impact of increasing the number of hard core unemployed in government training programs: (*a*) Programs were still "economically efficient." (*b*) There were greater gains by trainees with less than nine years' education over trainees with greater than nine years' education; therefore, the basic education portion of training is of most value. (*c*) Training was more beneficial for those less than 22 years of age than for those greater than 22 years of age. (*d*) Trainees gained financially after undergoing training.

S. Allison (1969) developed a cost-estimating model for undergraduate pilot training. Inputs to Allison's model consist of or can be (*a*) undergraduate pilot training graduation requirements, (*b*) course requirements, (*c*) instructor-student ratios, (*d*) administrative and support manpower relationships, (*e*) number of aircraft and simulators available, (*f*) quantity of facilities available, and (*g*) cost relationships. The model, given the inputs, computes the cost required for training in terms of research and development costs, investment costs, annual operating costs, and long-range feasibility estimates.

The Ozarks Regional Commission presented a rather detailed account of their cost-effectiveness system (Manuel, 1970). The goal of the commission is closing the "income gap" between the Ozark region and the rest of the nation. They wanted to measure the additional value of occupational education in the Ozark region. They saw their major problems as transposing the gains and losses into dollar terms. Benefits are calculated in terms of what buyers and users of the commodity will pay, or in terms of production costs if the former are not available. Costs consist of the value of the goods and services used up in the project as compared with their use for other purposes. This is called the value of alternate uses. If no alternate use exists, the costs are zero.

Intangible costs and benefits cannot be put into dollar terms, but they can be quantified and compared in terms of alternate courses of action. If, among two projects, *A* gives more net benefits than *B*, but if *B* has intangible benefits which override the net benefits of *A*, then *B* might be chosen as the course of action.

Some of the Ozark commission's cost-effectiveness formulae are presented:

1. **cost benefit of the program =**

$$\left(\frac{\text{program cost per student} + \text{tuition and books per student}}{\text{annual income generated per student}}\right) \times \left(\frac{\text{enrollees} - \text{dropouts}}{\text{enrollees}}\right) \times \left(\frac{\text{program length in months} - \text{student participation in program in months}}{12}\right)$$

2. **facility cost program =**

$$\left(\frac{\text{enrollees} - \text{dropouts}}{\text{enrollees}}\right) \times \left(\frac{\text{program length}}{12}\right) \times \left(\frac{\text{space allocation in square feet}}{}\right) \times \left(\frac{\text{space cost}}{\text{cost period}}\right) \times \left(\text{cost benefit of program}\right)$$

3. **cost benefit equipment =**

$$\left(\frac{1}{\text{enrollees}}\right) \times \left(\frac{\text{length of time equipment available}}{\text{time equipment used in months}}\right) \times \left(\frac{\text{equipment cost}}{\text{period equipment usable (10 years)}}\right) \times \left(\text{cost benefit of program}\right)$$

Gain Scores and Final Examination Grades

Carver (1966, 1969, 1970) presents a rather conclusive argument against the use of gain or difference scores in evaluation research. The problem in the before-and-after measurement of gain scores is that when small significant increases are registered, there may actually be a tremendously large increase in knowledge. This paradoxical result comes from the inequality of measurement at different points along the scale. Carver hypothesizes that a curvilinear relationship exists between test scores and knowledge, with knowledge increasing faster than test scores. One can rarely find a significant positive correlation between initial test scores and gain scores (often there is an inverse correlation). This is contrary to expectation, since it is expected that the more intelligent student will learn more and that the more interested student will be motivated to study more. One can partially explain this finding on the basis that students who already know a lot do not have much left to learn. Another related problem is the ceiling effect which occurs when the initially bright student already has most of the items on the pretest correct and does not have much room for improvement. Carver indicates that final examination grades constitute a dependent variable measure that is superior to gain scores, but with certain restrictions: The ratio of final knowledge to initial knowledge must be considerably greater than one; the correlation between initial knowledge and final knowledge must remain high; and the variance of final knowledge must be greater than the variance of initial knowledge.

Carver (1969) offers another solution—one involving separation of the initially bright from the initially dull students. This is done to correct a motivation problem for the initially high scoring student who has to waste time completing items at a low level. It is possible that if the bright student started off at a higher level, his gain may have been greater. On this basis Carver concludes that final scores are the best, because of unacceptable solutions using functions of initial and final scores and because expectations are not confirmed about initially bright students. Guilford (1970), though, feels that absolute scaling methodology might offer a solution to this dilemma.

Bereiter (1963) presents certain other related problems in the measurement of change:

1. The "overcorrection–undercorrection dilemma" which occurs when there is a negative correlation between the initial score and the gain score. This can be corrected so that a positive correlation can exist between initial and gain scores.
2. The "unreliability–invalidity dilemma" which occurs when there is a high correlation between pretest and posttest, thus lowering the reliability of the difference scores. If one obtains reliable difference scores because of a low pretest–posttest correlation, then the less we can say about the gain.
3. The "physicalism–subjectivism dilemma" which involves the choice of the scale units given versus units conforming to psychological meaningfulness. Bereiter recommends the use of terminal scores because change scores create too many problems.

Confidence Testing and Partial Knowledge

Shuford, Albert, and Massengill (1966) and Shuford (1967) have constructed a scheme to provide for more adequate measurement of student knowledge than is possible with traditional testing methods. They feel that additional information is available from the student's degree of belief probabilities. A mathematical system is presented which ensures that a student can maximize his expected score if he truly reflects his degree of belief or probability that a specific response choice is correct. With the traditional procedure, using a true-false test as an illustration, the student assigns a different probability for each response depending on his state of knowledge. If the student sees the probability of true as being greater than .50, he should choose true; but if the probability is less than .50, he should choose false; if it is equal to .50, he can choose either response. Generally, a student with poor knowledge (p = .51) will get the same score (if correct) as the person with good knowledge (p = .90); therefore, the choice situation loses data about the student's knowledge. In confidence testing, the student receives a confidence score (a function of probability) if his answer is correct plus a score for the correct answer. In addition, the student can receive credit if he is certain that his response is incorrect and the response is, in fact, incorrect. In one study (Massengill & Shuford, 1969), using multiple-choice tests, confidence was divided among the choices to total 1.0. The subjects for this study were 26 college-level students. It was

found that the confidence ratings were highly related to the probability of their answering the questions correctly.

Gardner (1970) administered a course pretest using confidence estimates to 151 student instructors. The test was designed to determine necessary training for these instructors. Even with the confidence scoring, there was no significant correlation of the pretest with practice teaching or with final class standing. The author still claims that confidence testing yields a better assessment of student knowledge, as well as higher reliability.

Coombs, Milholland, and Womer (1956) present another method of assessing additional student knowledge. Traditionally, in scoring a four-choice multiple-choice question, a subject is given a point for the correct answer and no points for a choice of any incorrect answer or distractor. Partial knowledge exists when the student can identify one or more of the distractors. Using this technique, in a multiple-choice format, one point is given for each distractor identified and three points are subtracted if the correct answer is identified as a distractor. Scores on each four-choice item can range from plus three to minus three. Partial-knowledge testing, then, yields increased item and test variance and penalizes for random guessing. Two possible disadvantages of this method are that it is not applicable to all kinds of tests (*e.g.*, true-false tests), and the scoring is time-consuming.

Characteristics of Material to be Learned

R. Allison (1960) gave 13 different learning tasks to 315 enlisted men at a United States Naval Training Center. Thirty-nine aptitude and achievement measures were also administered. Rate, curvature, and speed during the first and second half of the task were used as criteria of learning. Using factor analytic techniques, Allison found that learning was organized in a multidimensional way. Therefore, he contended that learning is not a single trait, but contains several factors depending ". . .upon the psychological process involved in the learning task and the content of the material to be learned (p. *iii*)." Also, the aptitude and achievement measures had much in common with the learning measures, demonstrating that the ability to apply knowledge and the acquiring of knowledge are very similar.

Naylor, Briggs, and Reed (1968) found that a primary task (tracking) is performed better in conjunction with a coherent or meaningful secondary task (monitoring) than in conjunction with a less meaningful or coherent task. Therefore, secondary task coherence can affect primary task performance in dual learning situations.

Weitz (1962, 1964) determined that with different difficulties of independent variables (*e.g.*, amount of information given in a training task), the maximal effect on transfer of training will occur either early or late during the trials. For easy information the maximal effect occurs early and for difficult information the maximal effect occurs late.

Underwood (1969, 1970) performed several learning experiments which demonstrate a breakdown of the total-time law which states that the amount learned is a function of total study time. Eleven experiments were performed, each varying the frequency of massed and distributed practice. The results showed that (*a*) recall of distributed practice was always greater than recall of massed practice; (*b*) massed practice words which were presented with the same exact frequency as distributed practice words were judged to have been presented less frequently; and (*c*) the difference (in recall) between massed and distributed practice increased as the frequency of repetition increased. Underwood hypothesizes that the difference between massed and distributed practice could be due to a failure of reception under massed practice which resulted in learning as if under a less frequent rate of presentation.

Jensen (1971) gave two groups of high school students equivalent forms of a visual and auditory digit span test. Both forms were administered to both groups in a counterbalanced order under immediate and 10-second delayed recall conditions. Jensen found that auditory memory was better than visual memory for immediate recall, but that the reverse was found for the 10-second delay condition.

Rather than viewing instruction as merely presentation of information, Whitmore (1970a) feels that it is a way of controlling student behavior so that learning takes place. Some factors which affect verbal learning are (*a*) attention span, (*b*) organization of the material into meaningful units, and (*c*) sequencing of material (*e.g.*, hierarchical, whole → part, and general → specific).

Carkhuff (1969) concluded relative to counsellor training that ". . . those programs in which high-level functioning trainers focus explicitly upon dimensions relevant to helper gains and make systematic employment of all significant sources of learning, including, in particular, modeling, are most effective (p. 244)."

Composition Scoring

Fostvedt (1965) constructed several criteria for the evaluation of high school English compositions in order to correct for non-uniformity of evaluation standards across teachers. Several sources were used to formulate the criteria: (*a*) coherence and logic, (*b*) development of ideas, (*c*) diction, (*d*) organization, and (*e*) emphasis. A sample of college English experts ($N = 9$) ranked these criteria. Kendall's coefficient of concordance was .75 ($p < .01$), indicating agreement among the experts as to the importance of each criterion. Next, 30 English teachers were asked to grade 20 themes as "above average," "average," or "below average" on each criterion. Analysis of variance was used to test criterion reliability, and the result was not statistically significant ($p > .05$); therefore, different teachers graded the same themes differently. Chi-square tests also demonstrated no agreement; hence, the criteria were not reliable when used for grading purposes.

Bushan and Ginther (1968) feel that there is a good deal of personal bias in grading essays and that a more objective method is needed. Differentiating between essays should take into account ". . . the structure and length of the sentence, vocabulary, and length as well as sociological and psychological construct of the test (p. 417)." A computer program was used which read off and quantified several relevant, scorable variables on 11 University of Chicago essays which were also graded by three experts. The three best and three worst essays were then coded for the computer and so analyzed. Thirteen criteria were employed to determine differences. After the differences were ascertained, these were used on the remaining five essays. Overall results demonstrated that better essay writers (*a*) have a larger vocabulary; (*b*) include statements of other authorities who are named; (*c*) give exact dates for events; (*d*) use numbers for quantities; and (*e*) use fewer words from psychological categories that can be analyzed for personality differences.

Testing

Much of the previous discussion in this chapter has been concerned with various applications of testing. In this section, testing in the pure sense is discussed.

Paper-and-pencil tests, as the name implies, are tests which the examinee takes with a printed test and a pencil. Most tests of this type require at least some reading ability. Some types of paper-and-pencil tests, though, require no reading ability at all. Many perceptual speed and perceptual motor tests are available on the market. Users of perceptual tests feel that they are related to some performance aspects of jobs. The verbal type of paper-and-pencil test should be used only in jobs which are primarily verbal or cognitive in content. It would probably be inappropriate to give a paper-and-pencil intelligence test or a vocabulary test to a person applying for a mechanical trade. Such tests, however, would be appropriate for some clerical positions. In performance tests (Danzig & Keenan, 1956; Fiske, 1954), the trainee or employee is asked to perform some tasks in which the content is relevant to his present or future job. Some performance tests are less obviously related to jobs than others. Performance tests can range from dominoes, mazes, and puzzles to performance of job tasks using real job equipment. Perhaps the most sophisticated type of performance test is the proving ground. In the proving ground (McSheehy, 1959), the trainee is placed on the job. An attempt is made to cycle him through all the job tasks in a short period of time. As he performs each task, the trainee is evaluated and he, in turn, evaluates the training in relation to the job.

Statistical Methods

There are a number of little used and less understood quantitative methods which can be useful for training evaluation and student achievement measurement.

Partial Correlation and Part Correlation

Partial correlation, according to DuBois (1957) is ". . . the Pearson product-moment correlation between two sets of residuals, from both of which variance associated with the same set of independent variates has been eliminated (p. 192)." In actual practice partial correlation is used to hold one or more extraneous or contaminating variables constant. For example, in calculating the correlation between height and weight, one might wish to hold age and sex constant. Part correlation, on the other hand, is "defined as the Pearson product-moment correlation between a set of residuals on one hand and an unmodified variable on the other. . . " In studies of learning, for example, it may be pertinent to inquire into the degree to which final standing in some skill, less the variance related with initial standing, is related to some outside predictor variable (p. 60)." The use of this statistic (part correlation) will help to clarify some of the problems associated with the use of raw gain scores mentioned earlier in this chapter.

Factor Analysis

Factor analysis is simply a statistical method for eliminating the redundancy present in correlation matrices. One might, for example, be able to reduce a 20 by 20 correlation matrix to a 20 by 5 factor matrix, thus using only five factors rather than 20 items to describe the matrix.

Obviously, factor analysis can be a useful tool in training evaluation and student achievement measurement. For example, one might have a 15-item rating scale which measures on-the-job behavior of training school graduates. It would be inappropriate to describe the on-the-job behavior of these men in terms of either 15 separate dimensions or one overall composite when the 15-item rating scale might be reduced to three or four dimensions which more parsimoniously describe on-the-job behavior. If predictor tests were used, then, significant validity coefficients might be dependent upon whether or not one used factor analysis. Bergman (1970) had such an experience when attempting to predict the behavior of 139 oil company salesmen.

Another old technique, but one which will probably be used more frequently during the next decade, is Q-factor analysis. In performing a Q-factor analysis, one simply factor analyzes the matrix of person correlations rather than item correlations. This method can be useful for grouping persons who think or behave similarly. For example, when constructing a training program, it may be useful to know the different cognitive styles of the potential trainees so that the training can be adapted to the needs of each homogeneous group. Eddy, Glad, and Wilkins (1967) used Q-factor analysis and found that their training program differentially affected ". . . .students depending upon their own goals, attitudes, and characteristics and of their work environments (p. 23)."

Tucker (1966) recently presented a rather unique application of factor analysis to the measurement of student learning. His innovation, though, has undeservedly been ignored by all but a few members of the behavioral science community. Using the Ekhart-Young theorem (a fundamental matrix decomposition theorem of factor analysis), Tucker found that individuals learn in qualitatively different ways over trials such that individuals can be grouped or clustered according to the way they perform or learn. Tucker would not use a single, homogeneous learning curve to describe what is, in fact, a heterogeneous phenomenon.

Canonical Correlation

Canonical correlation is an extension of factor analysis to the situation in which two separate sets of variables exist. The first canonical correlation is the highest correlation between a principal component of the first set of variables with a principal component of the second set of variables. The second canonical correlation is the correlation between a second principal component of the first set of variables with a second principal component of the second set of variables. Canonical correlations are continually extracted until all the common variance between both sets of variables is accounted for. The method is most applicable when there are two separate sets of variables: for example, one set of predictor variables and one set of criterion variables.

Moderator Variables

A test is a moderator when its score differentially determines the predictability of another test or variable. For example, one may be able to adequately predict the performance of college students using an intelligence test for those who score high on a test of achievement motivation, but not for those who score low on the test of achievement motivation. Race is one of the more currently popular moderator variables. Much recent research has shown that employment tests are differentially predictive across racial groups, thus supporting the contention that common selection standards for Negroes and whites are inappropriate or unfair. Moderator variables are sufficiently important to student achievement measurement and training evaluation that they are given separate treatment in another chapter of this review.

Convergent and Discriminant Validity

Campbell and Fiske (1959) would define convergent validity as a high correlation between tests purporting to measure the same thing, while discriminant validity would refer to independence of tests measuring different factors. The one criterion for convergent validity is that the correlations between several tests measuring one trait must be significantly greater than zero (mono-trait hetero-method correlation). For discriminant validity, three criteria must be met: (*a*) The single-trait-multimethod correlations must be significantly greater than the correlations not having trait or method in common; (*b*) the single-trait-multimethod correlation should be significantly higher than different traits measured by the same

method; and (*c*) there should be a stable pattern of trait interrelationship regardless of the method used.

Campbell and Fiske advocate the use of a multitrait-multimethod matrix which is in reality confusing and unnecessary, since all that is required is understanding of the concepts involved. Dielman and Wilson (1970) and Kavanagh, MacKinney, and Wolins (1971) are among those who have successfully applied this technique.

Internal and External Validity

Campbell and Stanley (1963) define internal validity as "significance," and external validity as measured change in job behavior. Campbell, Dunnette, Lawler, and Weick (1970) indicate that internal criteria are those that are directly tied to training behavior and that external criteria measure subsequent change in job behavior.

Campbell and Stanley (1963) and Winch and Campbell (1969) provide an exhaustive list of "threats" to internal and external validity. The threats to internal and external validity are (*a*) history or antecedents, (*b*) maturation of subjects, (*c*) testing effects, (*d*) instrumentation, (*e*) statistical regression (extreme scores), (*f*) differential selection of comparison groups, (*g*) experimental mortality, (*h*) selection-maturation interaction, (*i*) pretest sensitization, (*j*) interaction between selection bias and the experimental variable, (*k*) instability and unreliability of measures, (*l*) conditions making the experimental setting atypical or artificial, (*m*) multi-treatment interference, (*n*) irrelevant components of complex measures, (*o*) failure to replicate entire relevant parts of the experiment, (*p*) effects of experimental arrangements, and (*q*) effects of prior treatments. These writers recommend the use of experimental designs and statistical treatments which minimize the effects of these variables.

To assess effects of training, Campbell, Dunnette, Lawler, and Weick (1970) recommended using the following experimental paradigm:

Subject	Pre-measure	Training	Post-measure
Group I	Yes	Yes	Yes
Group II	Yes	Placebo	Yes
Group III	Yes	No	Yes
Group IV	No	No	Yes

In this design, the placebo group is necessary because the measureable effects of training can be attributed to the "Hawthorne effect." The post-test group (IV) is needed to avoid the possible effects of pretest sensitization.

Scaling Techniques

Siegel and Schultz (1960), Siegel, Schultz, and Benson (1960), and Schultz and Siegel (1961a, 1961b) report the use of scaled behavioral checklists to evaluate job performance in several Naval job specialties. These lists, developed on the basis of Thurstone and Guttman scaling principles, allow one to evaluate a man's proficiency by checking just one task on a list. If he can perform that task, it can be assumed that he can perform all tasks below that level on the scale.

Stone and Sinnett (1968) sought to determine whether or not the four-point grade point average distribution can be represented as being an equal interval scale. Thirty-six members of the University of North Dakota were used as judges. The grade range of A to F was divided into 12 intervals, *e.g.*, F to F^+, F^+ to D^-, D^- to D, D to D^+, . . . A^- to A. The judges were then asked to choose the grade intervals they thought were larger. They used the paired-comparison technique to rank all intervals. The median coefficient of consistency for all judges was .83. A scale was then constructed using Thurstone techniques. The results of this scaling analysis were that (*a*) the judged scale was found to be a logarithmic scale which could be compared to the grade point average scale; (*b*) generally, the intervals were judged to be smaller as the grade levels decreased; (*c*) the midpoint of the scale was between C^+ and B^-; (*d*) the distance between the midpoint of the grade to the (+) point appeared larger than from the (–) point to the midpoint; and (*e*) intervals containing a grade boundary were judged larger than those within a grade (*e.g.*, C^+ to B^- was thought greater than C to C^+).

Schultz and Siegel (1962a, 1962b) used multidimensional scaling analysis which integrates psychophysical judgments and factor analysis. The procedure is " . . . obtaining a matrix of interstimulus distances (psychophysical judgments) and . . . determining the dimensionality of the space containing the stimulus points (p. 3)." This method recognizes the multidimensionality, as opposed to the unidimensionality, of job performance criteria. Eighteen tasks performed by the avionics electronics technician were delineated. Judges were then required to indicate, along a scale, the distance or similarity between all possible pairs of tasks. After the analysis was completed, four job dimensions were found: (*a*) electro-comprehension, (*b*) equipment operation

and inspection, (*c*) electro-repair, and (*d*) electro-safety. Schultz and Siegel (1964) then used these four dimensions to construct unidimensional scales via Thurstone and Guttman techniques. Siegel and Schultz (1963) and Schultz and Siegel (1963) also applied multidimensional scaling analysis to classification of circuit types and to the Naval aviation electronics technician supervisor rating.

Signal Detection

Siegel and Pfeiffer (1969) and Siegel, Fischl, and Pfieffer (1968) were successfully able to apply signal detection theory to the prediction of academic success in both a military and a college setting. Signal detection theory " . . . provides a way of controlling and measuring the criterion the observer uses in making decisions about signal existence and provides a measure of the observer detection sensitivity (d') that is independent of his decision criterion (p. 145)." Eighteen subjects in Naval electronics training were divided into journeyman, intermediate, and advanced levels of training. Also, 40 male college sophomores were divided into high grade point average (2.88) and low grade point average (1.67) groups. The college sample was given a 49-item (psychology) true-false test, and the military sample was given a 23-item (circuitry) test. Items that are answered true are considered signal while items answered false are considered noise. A sensitive observer is one who differentiates with few errors between signal and noise. The results of this study were that (*a*) the d' was 2.16 for the high grade point average students, and 1.58 for the low grade point average students; (*b*) Naval technicians with the least training and experience had a d' of .64, while those with the most training and experience had a d' of 3.20; (*c*) analysis of variance results were significant for both groups at $p < .01$; (*d*) Scholastic Aptitude Test (SAT) scores were related to the college sample grade point averages; (*e*) other academic predictors did not correlate significantly with d', suggesting that it measures a different basic process; (*f*) SAT scores accounted for 16 percent of the high grade point average variance and 13 percent of the low grade point average variance; but with the addition of d', the predictable variance increased to 33 percent and 51 percent, respectively; and (*g*) the variance accounted for by the military tests was 11 percent, but it increased to 50 percent with the addition of d'. The authors conclude that d' can be used both as a predictor of performance and as a measure of training success.

The theory of signal detection bears an obvious relationship to the previously mentioned concept of confidence testing. Test scores based on confidence testing should correlate higher with signal detection variables (d') than with traditional test scores. Indeed, several investigators (Clarke, 1964; Pollack & Decker, 1964) have used confidence estimates in their signal detection studies. Signal detection, multidimensional scaling, and confidence testing all derive from experiments based upon psychophysical principles which are discussed in the next section.

Psychophysics

Siegel and Federman (1970) combined the magnitude estimation technique with peer group ratings to arrive at a novel method of performance evaluation. The subjects for this experiment (N = 20) were two groups of 10 avionics technicians. Each man was asked to estimate the number of uncommonly ineffective and uncommonly effective performances across nine performance dimensions for the nine other men over a specified period of time. The ratio of the number of uncommonly effective (UE) performances divided by the number of uncommonly effective performances plus the number of uncommonly ineffective (UI) performances ($\Sigma UE/EUE + \Sigma UI$) yields an index which varies between zero and one. One of the two groups was more experienced than the other, and this technique was able to differentiate between them.

In addition to the aforementioned study, Siegel and his associates at Applied Psychological Services have over the years applied the classical psychophysical methods to several other aspects of military and performance evaluation. Terminal threshold concepts were applied to electronics troubleshooting performance evaluation (Siegel, 1968). Psychophysical methods were used to maximize the probability of operator malfunction recognition (Miehle & Siegel, 1967). Activity circuit interactions were related to perceived circuit complexity (Pfeiffer & Siegel, 1967b). Magnitude estimation and the structure of intellect model were used to relate electronics maintenance job activities and the intellective scale values of these activities (Pfeiffer & Siegel, 1967a). The psychological relationship between perceived circuit complexity and a physical measure of circuit complexity was ascertained (Pfeiffer & Siegel, 1966). Magnitude estimates of perceived circuit complexity were related to subjective and objective job correlates (Siegel & Pfeiffer, 1966b). Magnitude estimation was used to measure avionics maintenance personnel subsystem reliability (Siegel & Pfeiffer, 1966a). And, finally,

magnitude and category psychophysical scaling methods were used by journeyman electronics personnel to scale the complexity of various aspects of their own jobs (Pfeiffer & Siegel, 1965).

Summary

The first section of this chapter presented an overview of some of the kinds and characteristics of dependent measures used in training evaluation and student achievement measurement. The test construction portion of this chapter contained a brief discussion of the steps to be followed in constructing a test plus some studies using novel tests or testing techniques. Other topics reviewed in this chapter were (*a*) hierarchical and sequential testing, (*b*) criterion- and norm-referenced testing, (*c*) performance evaluation problems, (*d*) cost effectiveness, (*e*) gain scores and final examination grades, (*f*) confidence testing and partial knowledge, (*g*) characteristics of the material to be learned, (*h*) composition scoring, and (*i*) statistical methods.

IV. LEARNING STYLES AND MODERATOR VARIABLES

Scope of the Problem

The sensitivity and predictive power of student measurement and training evaluation techniques can often be increased through the use of moderator variables. This is because certain attributes of select groups tend to make the testing evaluation methods more or less appropriate for the groups. Some of the factors which can be used as moderators are (*a*) achievement level, (*b*) personal and environmental variables, (*c*) social background factors, (*d*) cognitive style, and (*e*) affective reactions.

Cognitive styles are modes of thought, perception, and memory; they are also information processing habits. Some of the various types of cognitive styles that have been identified are (*a*) field dependence-independence, (*b*) attention span (or span of awareness), (*c*) breadth of categorizing (*e.g.*, lumpers and splitters), (*d*) conceptual styles (*e.g.*, modes of categorization), (*e*) complexity versus simplicity in word perception, (*f*) reflective-impulsive, (*g*) leveling versus sharpening, (*h*) susceptability to cognitive interference, and (*i*) ability to accept unrealistic experiences. French (1963), using a factor analytic approach, delineated two types of problem solvers: (*a*) those using a systematizing approach and (*b*) those using a scanning approach.

Rundquist (1969) contends that item analysis, factor analysis, and moderator variables have not helped to increase predictive efficiency because these various methods fail to take into account the fact that different antecedents can produce the same behavior across individuals (*e.g.*, visual recall via eidetic imagery or by short term memory). According to Rundquist, one must learn the mediating processes used by individuals in learning to do a job and then construct tests for the antecedent behaviors. These new tests would be better measures of an ability than more global tests, and they could avoid confounding effects. The new test or measure may be slanted more toward one antecedent than another, thus increasing the validity coefficient.

The overall trend towards individualization has caused some writers (Whitla, 1969) to plead for more research on student types, class mix, and the disadvantaged. Others (Bligh, 1965) have called for increased differentiation of norms for different groups (*e.g.*, sex, race, locale). Finally, some others (Project Impact, 1970) claim that computer assisted instruction and other forms of individualized instruction are the best way to account for broad student differences.

On the debit side, Gagne (1968) disputes the existence of learning styles. He thinks computer assisted instruction puts too much stress on the machine rather than on the student. He does, though, emphasize the need for individualized instruction, and he acknowledges the idiosyncratic nature of the student. Cohen (1970) feels that one must be careful when using cognitive styles as moderators and instructional aids, since they can change over time. For example, much of Piaget's work has shown that the child's problem solving style and conceptual mode of thinking will qualitatively change from infancy to adulthood. Cohen concludes that a valid decision about an individual's cognitive style at one time may prove to be invalid at another time.

One final note concerns the special case of the moderator variable approach when aptitudes or aptitude test scores interact. When this occurs, differential treatment of groups is mandatory. If not, erroneous or contaminated results will occur.

Motivation and Types of Intelligence

There has been a plethora of recent research emphasizing the effects of differential motivation and differential thinking styles (erroneously termed "intelligence") on student achievement.

These concepts certainly should be held in mind by anyone concerned with student achievement, from either the measurement or the instructional point of view. However, the payoff of the studies in these areas seems, as yet, indeterminate and problematical. Many of the studies are contradictory in results, and others require cross validation before their indications can be fully exploited.

Jensen (1969) postulates that there are two types of intelligence, abstract and associative, and that instruction and testing should be differentially tailored to suit these different modes of learning.

Rimland (1969) also suggests that there are two types of intelligence, practical and abstract. Rimland hypothesizes that practical intelligence is needed for job performance, and that abstract intelligence is needed for academic work. Such thinking would imply that most trade schools should rely heavily on job performance testing to measure student achievement. Rimland says that the traditional *g*, or general intelligence factor, measures "intracerebral events," or the ability to abstractly manipulate symbols and events in the head. This is the ability required of test takers. Others are better at "extracerebral events," or the ability to sustain attention on and perform simple tasks which simulate the job (*e.g.*, perceptual speed). Rimland posits that these two types of intelligence are mutually exclusive. In his research, he found that intelligence test scores correlated much higher with school grades than did performance test scores, but that performance test scores correlated much higher with job performance than did intelligence test scores. He concludes that different types of training and separate types of measurement are needed for students with different types of intelligence.

Rotter (1966) conceives the effect of reinforcement on behavior as dependent on whether the person perceives a causal relationship between his own behavior and the reward. If not, the result is attributed to luck or to the control of others. *Internal control* exists when the student thinks reinforcement is contingent upon his own behavior, while *external control* is when the student thinks reinforcement is controlled by others or by chance events.

In one study investigating the internal-external control thesis (Scott & Phelan, 1969), three groups of hard core unemployables were tested with Rotter's Internal-External Control Scale. The subjects in all three groups were matched on age, socioeconomic status, and scholastic aptitudes. The results demonstrated that black and Mexican American subjects demonstrated greater external control than did white subjects. The authors concluded that the externally controlled subjects did not feel that there was a relationship between individual effort and reward; therefore, they did not work unless given external reinforcement (*e.g.*, praise, money).

Atkinson (1966) presents a somewhat more vigorous theory of motivation involving achievement motivation, incentive, and goal expectancy. Atkinson's theory is depicted by the formula:

Motivation = f(motive x expectancy x incentive)

With motivation to approach a goal (nAch) held constant at 1.00 and with expectancy and incentive equal to .5, then the probability of goal approach is .25 (the highest possible). Atkinson defines incentive as the goal attractiveness, and motive as the ability to strive for satisfaction or to accomplish. "The strength of motivation to approach decreases as probability of success increases from .50 to near certainty (p_s = .90), and it also decreases as p_s decreases from .50 to certainty of failure (p_s = .10) (p. 17)."

From this formulation, it is easily seen that the young, deprived black child will rarely encounter a probability of success of .5 or greater. Because he perceives a certainty of failure, he then lacks the motivation to approach a goal; therefore, he does not perform as well in student measurement situations as the non-deprived white child who perceives a higher probability of success.

Katz (1967) more or less integrates the two earlier theories into a coherent two-stage theory of development which possesses implications for student measurement. During the first stage (up to two years of age) of development, the child's verbal efforts are normally reinforced by parental approval. Selective approval, on the part of the parents, can develop strong habits of striving for proficiency in the child. During the second stage, the parental standards and values of achievement are internalized by the child. "The child's own implicit verbal responses acquire through repeated association with the overt responses of the parents, the same power to guide and reinforce the child's own achievement behaviors . . . . Internalization doesn't take place until strong externally reinforced achieving habits have developed (p. 5)." Lower class children (including most blacks) are more dependent upon others for social reinforcement in academic situations. Lacking internalization, they will avoid achievement situations and

concentrate on other situations regarded as more promising. "Lower class Negro children tend to be externally oriented in situations that demand performance. That is, they are likely to be highly dependent on the immediate environment for the setting of standards and the dispensing of rewards (p. 8)."

Hess and Shipman (1965) present a very interesting and alternative developmental formulation. They feel that cognitive growth is " . . . fostered in family control systems which offer and permit a wide range of alternatives of action and thought and that such growth is constricted by systems of control which offer predetermined solutions and few alternatives for consideration and choice (p. 870)." In the deprived family context, the parent-child control system " . . . restricts the number and kind of alternatives for action and thought that are opened to the child; such constriction precludes a tendency for the child to reflect, to consider and choose among alternatives for speech and action. It develops modes for dealing with stimuli and with problems which are implusive rather than reflective, which deal with the immediate rather than the future, and which are disconnected rather than sequential (pp. 870-871)." Hess and Shipman performed a research study using deprived (black) and non-deprived mother and child pairs which supported their hypotheses. These authors concluded that the family shapes the modes of communication in the child, which in turn shape his thought and problem solving style.

In summation, these four positions suggest that, in both curriculum development and student measurement, differences in cognitive style and motivation must be accounted for in any program which purports to be at all comprehensive.

Race and Aptitude as Moderator Variables

In a recent survey of 13 studies, Boehm (1971) found that job knowledge and performance test criteria always yielded the highest validities. Generally, there are fewer validity differences between racial groups when these more objective criteria are used instead of ratings or rankings.

McFann (1969a, 1969b) noted that the differences between high- and low-aptitude men in Basic Combat Training were greatest on cognitive tasks, and that the differences were not as marked on motor skills and proficiency tests. In a project SPECTRUM study, high-, middle-, and low-aptitude groups were selected, and individualized training was given using videotape, one-to-one student-teacher ratio, feedback, reinforcement, and small increments. In some tasks, low-aptitude men reached standard but took two to four times longer; in other cases they did not master the material at all. McFann also found that high-aptitude groups learned equally well with lecture or individualized training, while low-aptitude groups learned well with individualized training, but not with lecture.

Foley (1971) wanted to determine if the Officer Qualification Test (OQT) was biased against blacks in determining final Officer Candidate School (OCS) grade point averages. The final OCS grades of blacks from caucasian colleges were not significantly different from a matched white sample. Blacks from Negro colleges, though, did receive significantly different grades than their matched white subjects ($p < .005$). In general, the OQT predicted better for the white sample, even though it was significant for both races.

Guinn, Tupes, and Alley (1970a, 1970b) wished to determine if the prediction of training success varied across subgroups. If this is the case, then overall predictive efficiency suffers. These writers found differences in training performance across race, area of the country, and education. All three differences, though, were not found in all occupational specialties. It can be inferred from these results that factors such as race and variations in cultural opportunity, as may exist across different educational and regional groups, can account for the differences in test scores across groups.

In a study performed at the American Telephone and Telegraph Company (Grant & Bray, 1970), task proficiency after training was used as a criterion because the investigators thought that it was uninfluenced by supervisory bias, peer pressure to control output, and motivation.

Five hundred subjects, both blacks and whites, who met and failed to meet normal selection standards were involved. Seven hierarchical levels of training were employed using tasks regularly performed by craftsmen. Pretest and posttest tasks were given at each level, and the highest level completed was the criterion. The results demonstrated that all selection instruments correlated with highest level passed, and there were no differences in minority and non-minority correlations. The School and College Abilities Test plus a test of abstract reasoning yielded a multiple R of .49 when correlated with the training criterion.

Age and Sex as Moderators

Using the Gates Reading Readiness Test and the Metropolitan Achievement Test for elementary school students, Miller and Norris (1967) found that younger school entrants were at a disadvantage at the start. This effect, though, disappeared after the first grade. The late entering group tended to have more achievement and psychological referral problems than the early and normal entrant group.

Gay (1969) investigated the differential effectiveness for males and females of three computer assisted instruction (CAI) treatments on delayed retention of mathematical concepts. The three methods of presentation were (*a*) "variable example" which depends on the subject's pre-instruction retention index as measured by the Gay Retention Index; (*b*) "choice" which allows the subject to decide on how many examples he needs; and (*c*) "fixed" which allows the subject three trials per mathematical concept. Fifty-three eighth grade subjects (27 male and 26 female) were randomly assigned to the treatments. The results indicated that (*a*) the females in the variable example group performed better than the females in the fixed and choice example groups ($p < .05$); (*b*) males in the choice group performed significantly better than females in the choice group ($p < .05$); (*c*) males in the choice group performed significantly better than males in the variable example and fixed groups ($p < .05$); and (*d*) females in the variable example group performed better than males in the variable example and fixed groups. Gay concluded that the choice method is best for males. Even though the males averaged three choices, they gave more trials to the difficult items and fewer trials to the easier items. The Gay Retention Index, though, seemed to be good for selecting the number of items for females.

Cross-National Evaluation

Husen (1969) discusses cross-national evaluation and points out that such evaluations can be confounded because of a difference in objectives, which are different across boundaries, including different traditions, emphasis, age levels of introduction, and opportunity. Husen also points out that the real purpose of cross-national evaluation is ". . . not to make overall comparisons between countries – we are not engaged in an international contest – but to obtain meaningful comprehensive measures of both cognitive and non-cognitive outcomes and to relate these to a comprehensive set of input variables, including those which measure opportunity. Thereby, provisions are made for a fruitful multivariate analysis of how outcomes are related to inputs (p. 343)."

Summary

This chapter was concerned with the various effects of learning styles and moderator variables. First, moderator variables were defined and discussed. Following this was a presentation of several motivational and developmental theories which purport to lend some insight into how moderator effects materialize. Additional sections of the chapter contained studies of race and aptitude levels as moderator variables; age and sex as moderators; and problems of cross-national evaluation. It was noted that although the moderator variable approach appears to possess merit, moderators are often elusive. Their identification and their desirability may be dependent on a host of interactive effects. Thus, although no advanced program will ignore moderators, one should not anticipate that they will provide a pat solution to prediction problems.

V. CURRENT TRENDS

Trends

About ten years ago, Schultz and Siegel (1961a) perceived a trend in evaluation research which has since been demonstrated. They found that rather than investigating an overall performance criterion, it is better to use factor analysis or multidimensional scaling techniques to identify the important components of the job or training task. In the past, there has been too heavy a reliance placed on the single composite criterion. This practice is wasteful and hides useful information. More and more recent research has demonstrated that one score cannot possibly represent the multidimensional and orthogonal aspects of performance. Once the investigator arrives at multiple criteria, he can use a weighted sum of the subcriteria to arrive at a composite evaluation. Schultz and Siegel also stressed in the validation of training programs the need to determine if performance changes over time. If so, one might wish to sample performance at different times or determine if a longer time span is needed.

Merrifield (1965) agrees with Schultz and Siegel (1961a) about the need for more multivariate training evaluative studies. He places special emphasis in this regard on the special abilities student.

A second trend has been noted in terms of emphasis on cross-cultural training. Brislin (1970) presents a rather acid critique of most military cross-cultural training programs. The aim of cross-cultural programs, according to Brislin, is to allow the military to function behaviorally and effectively in a foreign environment. Most programs, though, do not have data on effectiveness, and the evaluative methods used are inadequate. When evaluations were conducted, they were too dependent on verbal and written reports of the trainees. More data need to be collected on the actual overseas behavior of trainees; therefore, responses to attitudinal questionnaires need to be verified by other means. Evaluation needs to be conducted by reseachers not associated with the program. Also, the attitudes of foreign nationals should be sampled. Techniques should be available to assess transfer of training to the actual foreign situation with more replication and followup training.

Fiedler, Mitchell, and Triandis (1970) and Worchel and Mitchell (1970) have recently described an exciting new technique known as the Cultural Assimilator, which is based upon the critical incident technique. In this technique, critical incidents are obtained in which the norms or behaviors across cultures are quite different. Questions are asked about the incident with multiple-choice answers and immediate feedback. A target sample from the host culture selects the correct multiple-choice responses.

An experiment recently performed by the Navy compared two- and six-week Vietnamese language courses. The results demonstrated that (*a*) graduates of either course met most objectives in that they were able to acquire some vocabulary and conversational skills; (*b*) students of higher aptitude performed extremely well in the six-week course; (*c*) the language laboratory produced problems which were later rectified; (*d*) many graduates thought the course was inefficient and that they did not use all that they were taught; and (*e*) low-aptitude students were only marginally adequate.

Predictive Evaluation

Richards, Holland, and Lutz (1967) found that non-academic accomplishment was relatively independent of academic achievement in college. Non-academic accomplishment in high school correlated .39 with non-academic accomplishment in college. On the other hand, the American College Testing Program's College Admissions Test correlated .29 with college grades, and high school grades correlated .38 with grades in college. The authors concluded that this study is important for college admissions officers who are interested in the non-academic as well as the academic potential of the students they accept.

Ryan (1968) compared students taking a conventional 12th grade mathematics course with students taking an experimental methematics course to determine if prior courses in high school can moderate performance in college courses. The students were also given a mathematics achievement test, a mathematics proficiency test, and a verbal ability test. The results showed that the mathematics achievement test correlated more highly with grades than did the mathematics proficiency test for the experimental group and visa versa for the the conventional group. Also, students in the experimental group performed significantly better than conventional students on mathematics achievement, but no better on mathematics proficiency or verbal ability. Hence, the achievement test probably reflects differences in prior instruction rather than differences in more general abilities.

Goolsby, Frary, and Lasco (1968) compared the results of the Florida Bar Examination with grades and aptitude test scores to determine if these latter measures could be used instead of part or all of the lengthy and expensive Bar examination. Only low correlations were found, causing the authors to conclude that no aptitude test scores or grades could supplant the Bar examination. In another law predictive context (Klein & Evans, 1968), nine experimental measures were correlated with law school success for 978 law students across several schools. Undergraduate grade point average turned out to be the best predictor of law school grade point average in some schools, while the Law School Admissions Test was the best predictor in other schools. The authors concluded that undergraduate achievement can predict graduate achievement for law school students. In another law school situation (Lunneborg & Lunneborg, 1967), 557 law school students were surveyed in order to ascertain which types of undergraduate courses predict law school success. Verbal, accounting, and language courses were found to be the poorest predictors, while philosophy, economics, history, and business administration were the best.

Kaplan, Freedman, and Kaplan (1968) wished to examine the utility of replacing clinical ratings of psychiatry students with the National Board of Medical Examiners Test. This latter test was found to correlate .44 with the ratings. These writers,

though, indicate that other types of information, in addition to the test score, are needed because the written examination does not account for enough of the variance of the dimensions being investigated by the ratings. The dimensions of personality and psychopathology are not assessed by the test, but they are assessed by the ratings. Some further investigation of the ratings seems warranted, though, since they are so much more subject to bias and error than tests.

Bergstrom (1968) related measures of school achievement to important job behaviors in order to evaluate a school curriculum. A sample of students (N = 150) was taken from three types of schools: (*a*) urban vocational, (*b*) urban comprehensive, and (*c*) suburban comprehensive. The results indicated that vocational training should stress personal adequacy and communication skills. The results of this study showed that (*a*) those employees with specific vocational training were more likely to be placed on a related job; (*b*) students with low grades (D) in vocational courses obtained lower job evaluation only in skill areas of the job; (*c*) graduates who were poor in school attendance tended to get significantly lower ratings; and (*d*) one-half of all trained workers were not placed or retained in a job they were trained for.

Bale, Rickus, and Ambler (1970) wished to determine if undergraduate aviation training could be used as a predictor of graduate or replacement air group (RAG) instruction. The traditional criterion for student aviators has been successful completion of undergraduate flight training, but this was felt inadequate because it did not account for RAG instruction. The grades in training were based on (*a*) air to air weapons, (*b*) air to ground weapons, (*c*) basic ground, and (*d*) instrument navigation. The multiple regression coefficient between training grades and success-failure in RAG was .43; in a cross-validation sample it was .36. Use of these prediction measures would have reduced attrition in RAG by 34 percent. The investigators also found that 15 tests gave a multiple R of .43, while four tests gave a multiple R of .38.

A final study demonstrates that OCS grades can be used to predict officer effectiveness (Rhea, 1965). The fitness reports of 2,183 OCS graduates were obtained after 18 months of service. A low, but significant, correlation between each OCS variable and fitness was obtained (average r = .22). In general, fleet fitness reports were less predictable than shore fitness reports. The best predictors were final school grades and military aptitude which had correlations ranging from .16 to .37.

Sensitivity Training

Another comparatively recent innovation involves sensitivity training and its associated methods including T-groups, role playing, and the like. Bass, Thiagarajan, and Ryterband (1968) are severely critical of sensitivity, or T-group, training. They say that ". . . we still may hear complaints about the lack of evaluation of sensitivity training, yet a bibliography of at least 50 evaluative studies now exists. . . . why have these studies failed to impress social scientists? . . . A major reason may be because insufficient attention has been devoted to the purposes of the evaluation and the public for whom the evaluation is being prepared" (p. 2f).

One very controversial study by Golembiewski and Carrigan (1970) involved an assessment of change resulting from sen 'ivity training. The sample in this study was 16 commercial sales managers. Progress was measured by self-report on the 48 items of Likert's (1967) Profile of Organizational Characteristics. The participants rated their organization twice, once as their conception of the ideal, and once as they perceived it to be in actuality. This was done both early in the week of training and four months after training. Both "ideal" and "now" scores increased in the interim in the "participative" direction, thus supporting the authors' hypothesis. The authors themselves acknowledge the possibility of the Hawthorne effect or other methodological weaknesses in their design, but tend to minimize such possibility in favor of true change. Becker (1970), though, seems to think the study is of little value for several reasons: Golembiewski and Carrigan failed to rule out alternative explanations; they indicated that the Hawthorne effect cannot be rejected, yet they rejected it; and they failed to account for changes which could have occurred through passage of time. Becker closes with ". . . changes did and probably continued to occur, so it may be permissible to sell such a design to managements; but under no circumstances should one attempt to sell such a design as science (p. 96)."

In another study (Cook, Hahn, & Sheppard, 1971), 23 Navy Medical Service officers took part in a three and one-half day management style seminar; a six-month intervening period at a duty station followed; then a two and one-half day management style session was conducted. In their training sessions, the officers were presented with

(*a*) problem analysis using "force field method;" (*b*) group ranking which allowed for cross-subject influencing; and (*c*) small group management style sessions. In the six-month intervening period, the subjects were urged to use their newly acquired techniques. The final session included discussion, reinforcement, and feedback of management style data. The Management Value Index (MVI), an index of management style, was given at the beginning and end of the first session, and at the end of the second session. The results indicated course influence. The Leadership Opinion Questionnaire was also administered, and the results indicated a decrease in structure without a corresponding decrease in consideration. These results are somewhat suspect, since participants thought their management styles were more open than did their colleagues and subordinates, especially with regard to participation. The authors concluded that the much larger value change between the second and third administration of the MVI suggests the need for an on-the-job "incubation period" in order for attitudes to change.

Federman and Siegel (1965), in a group dynamics study, isolated four performance-related communication factors from training teams in a helicopter simulator. These four factors were derived from a factor analysis of 14 communication predictors shown to be related to miss distance in antisubmarine warfare. The four factors were (*a*) probabilistic structure, (*b*) evaluative interchange, (*c*) hypothesis formulation, and (*d*) leadership control. In a second study, Siegel and Federman (1969) cross-validated the factors and developed a training course based on the derived factors. The trained group was found to perform better than a control (untrained) group in two performance tests involving enemy submarine detection and destruction.

Programmed Instruction

Lumsdaine (1970) feels that the most important contribution of programmed instruction is not improvement in instruction, but rather in the implicit requirement for clearly stated objectives in behavioral terms.

Mager (1970a, 1970b) maintained that it is impossible for the instructor to apply all the principles of learning in the classroom. This is not because he does not want to, but because the learning environment is prohibitive. "We still put large groups of students in front of a single instructor and insist that they all learn at the same rate (p. 4)." This procedure may be convenient and inexpensive, but it is inefficient. Programmed learning devices and machines are held to possess the potential for solving these problems since they usually (*a*) present instruction in small steps; (*b*) reinforce the student along the way; (*c*) help the student proceed at his own pace; and (*d*) feed back responses into the device to modify instruction to fit the particular needs of the student.

In sequential programming, learning proceeds in very small steps, and all learners go through the same steps. In alternate programming, though, the student's steps can be different, and they are governed by the student's own responses.

Keller (1968) indicated that the techniques of programmed instruction can be used in any classroom situation. However, according to Keller, one criterion that the instruction must meet is that it be individualized. Another requirement is that criterion-referenced testing be used.

Lindvall and Cox (1969) present a Structured Curriculum Model (SCM) for developing a programmed instructional course. They state that one must define specific objectives and organize them according to difficulty or prerequisites. This organization provides a structural sequence which is a frame for determining the student's present status and for his future planning. In the SCM, the curriculum materials must be matched to the objectives, and one must keep in mind that students can master the same objectives with different kinds of material. In addition, the student must be given a diagnostic evaluation to place him in the proper location along the learning continuum. The placement test should ". . . select items which test representive objectives along the continuum (p. 170)." Pretests are also suggested prior to each instructional unit, because the student may be able to cope with some of the objectives in the unit, and not others. Evaluation in this model is by way of "curriculum embedded tests" and "post-unit" tests. Curriculum embedded tests (*a*) measure one objective of a unit; (*b*) they are content-referenced; (*c*) they are short; and (*d*) they enable the teachers to make decisions regarding student advancement. Post-unit tests help the teacher to decide whether the pupil should progress to the next unit or should be given remedial work.

Glaser (1967) insists that uniformity within any one grade level can never be achieved because of individual differences. This results in the need for programmed or computer oriented instruction.

Glaser also suggests that too much research has been done comparing methods and not enough research has been done on learning what and how variables affect students. Glaser describes the requirements for individualized instruction that have been set forth at the Learning Research and Development Center:

1. Time limits and grade levels must be redesigned so the student works at his actual achievement level, and he progresses only after he has mastered the prerequisites for the next higher level.
2. Sequences of progression must be assigned to each student.
3. Progress must be continually assessed to modify the teaching program to fit pupil needs.
4. Materials should be provided to the student which will self-direct his learning.
5. Performance standards (feedback) should be provided to the student.
6. A data processing system should be provided so that the teacher can take advantage of detailed information about each student, and construct an appropriate program for him.
7. Pretests and posttests should be provided for each instructional unit.
8. Sequential testing procedures should be employed for initial placement.

Whitmore (1970c, pp. 33-34) recites four learning prinicples that are contained in automated individualized instruction that are not generally found in traditional instruction. These learning principles are (*a*) continuous participation by the student in the instructional process; (*b*) providing immediate knowledge of the results to the student for each response that he makes; (*c*) recognition of individual differences in rate of learning; and (*d*) providing a high rate of success for the student throughout learning.

The last principle, Whitmore says, is the most difficult to implement, since it requires very careful analysis of the material to be learned.

McFann (1969a, 1969b) characterizes training strategies and their characteristics as follows:

Strategy	Curriculum	Time	Standard
1	Fixed	Fixed	Variable
2	Fixed	Variable	Fixed or variable
3	Variable	Fixed	Variable
4	Variable	Variable	Fixed or variable

In this scheme, a fixed standard means that the student is to reach a minimal level, while a variable standard means that the student can go beyond the minimal level to another higher level.

Strategy 1 is only recommended when the input to the course is homogeneous; if it is not, there will be variable output. It ignores individual differences and involves the additional problem of where to set the level of training. Strategy 2 is similar to most present training in the military. Those who fail to pass the first time are recycled (variable output time). One can gear the training to low-aptitude men, or allow the more intelligent men to go through the program faster. Strategy 3 has a fixed time limit and will result in variable output. Strategy 4 is the most flexible and the most individualized, but it requires the best management.

Computer Assisted Instruction (CAI) and Testing

Computer assisted instruction represents one of the most recent innovations in training methodology. One of the main problems of CAI is its cost when compared with other similar methods which might give equivalent results (*e.g.*, TV) Another, more serious, objection to CAI is that it does not allow the student enough opportunity or freedom to chart his own progress (Hammel, 1969).

Hansen, Hedl, and O'Neal (1971) feel that computer assisted testing will come into full flower this next decade. One reason given for this is the evidence that people answer questionnaires more honestly when they are presented via computer than by traditional methods.

Holtzman(1971) says, "In a traditional setting, the instructor keeps a record of how well each student does on each achievement test for the course, while the periodically collected scores from standardized normative tests are stored centrally. When instruction is individualized, testing must be done more frequently and at different times for each student (pp. 547-548)."

Seidel (1969) discusses the purposes of project IMPACT which is to provide the Army with an appropriate and efficient CAI system adaptable to the individual trainee. Programs are to be branched and adapted to the entry characteristics of the trainee and his performance throughout instruction. Some of the important decision factors involved are (*a*) entry characteristics, (*b*) education and background, (*c*) responses of trainee, (*d*) response latency, (*e*) pattern and history of errors, (*f*) relation of individual and group norms to responses, and (*g*) subject matter.

Gagne (1968) disagrees with most of these writers regarding the usefulness of computers in testing (and instruction). He thinks that CAI puts too much stress on the machine rather than on the student.

Atkinson (1967) discusses three levels of CAI:

1. Simple - "fixed, linear sequence of problems (p. 56)." There is no method of changing the instruction as a consequence of the student's responses. They are also called "drill and practice" systems.
2. Complex - also called "dialogue" systems. They provide high-level interaction between student and system. The students can give many variations of response, can ask a variety of questions, and can generally control the sequence of learning.
3. Tutorial - are between simple and complex with regard to the student's interaction with the system. There can be decision making or branching, depending upon the student's responses. The students can, therefore, follow separate paths. One of Atkinson's findings was that fast learners, on a month by month basis, showed a continual improvement in rate of progress, while medium and slow students had constant rates of improvement.

Ferguson (1970) described how computer assisted criterion-referenced measurement was applied to an experimental school in individually prescribed instruction (IPI). Addition and subtraction skills were taught in a sequence in which each stage built onto and was required for the next stage. After each answer, the computer made a decision, on the basis of percentage correct and number of problems of this type attempted, whether to go to the next level or continue presenting problems of the same type. Each item was randomly selected from a population of similar items. Direct manipulation of type I or type II errors was possible. The type I error allows the student to progress to the next level prior to mastery; therefore, this is considered the most serious type of error.

Applications of Programmed Instruction

Yeager and Kissel (1969) hypothesized that the number of days needed to master a unit of instruction is related to the students' "initial entering state." The entering state variables were (*a*) unit pretest score which, when subtracted from 100, gives the distance or amount to be learned; (*b*) number of types of pretest skills on which the student failed to show mastery (IPI only concentrates on these); (*c*) intelligence; and (*d*) age which reflects student maturity. The entering state variables used in this study, therefore, were pretest scores, number of skills to be mastered, I.Q., age, and total units mastered previously. The results demonstrated that pretest score, numbers of skills to be mastered, and age were the best predictors, while I.Q. score had the least influence. The multiple correlation coefficients for different types of materials ranged from .65 to .84 (N=40).

Atkinson (1967) found that students in an experimental CAI reading program performed significantly better in all aspects of reading (*e.g.*, pronunciation, vocabulary, recognition) than did students in conventional (control) reading classes. The control group received CAI mathematics instruction, but not CAI reading instruction.

K. Johnson (1968) examined the results of three different methods of teaching military communications courses. The three methods used were conventional, programmed instructional booklets, and partially individualized (first week conventional followed by self-paced). The results showed that the self-paced (partially individualized) instruction produced a 16 percent reduction in course length, while the programmed instruction produced a 9 percent decrease in course length. These reductions were accomplished without loss of skill.

Geisert (1970) wished to examine the contribution of format and feedback to learning. Two groups of Army National Guardsmen (N=44) were used as subjects. All concepts to be learned in the experimental group were arranged hierarchically (mapped) to ease positive transfer to the next highest level. Fifteen dependent variables were used including reading time on booklet, test scores, time spent reading instructions, time spent on practice, and time spent on problem solving instructions. The results demonstrated no significant differences between the hierarchical group and the traditional group, except that the former group tended to do all things slightly faster. Similar results were obtained for the feedback-no feedback group. With regard to certain attitude scales which were administered, it was shown that subjects preferred to learn from the mapped-feedback system over the traditional system. The subjects also thought that a computer assisted screen was an effective way to present material when compared to booklet material, although neither was shown to be more or less effective than the other.

A novel and interesting approach to self-paced instruction was recently developed by Sheppard and MacDermot (1970). Subjects were 203 students enrolled in an experimental course and 98 students enrolled in a traditional course. The students in the experimental group were to study one of 36 sections of a psychology book. After study, the students were asked to explain the lesson in detail to another student who had already completed the work, or to an instructor. If the learner failed, he would repeat the lesson until mastery was achieved. Completion of all 36 interviews earned a grade of A, 75 percent a grade of B, 50 percent a C, and 33 percent a D. The control group was as comparable as possible, since the students spoke in small groups and used the same book. At course completion, both groups were given 100 multiple-choice questions and five essay questions. The control group was told that the final examination contributed 50 percent of their grade, while the experimental group was told that the final examination did not count. In addition, the control group was informed that they had to finish the entire test. These last two factors should produce a bias in favor of the control group. The mean for the experimental group on the multiple-choice test was 73.1, and for the control group it was 66.8 ($p < .01$). On the essay questions, the experimental group scored 17.4, and the control group 13.9 ($p < .01$). Also, composite student satisfaction, as measured by an attitude scale, was higher for the experimental group ($p < .01$). Of those queried, 94 percent thought the interview method was more effective than the lecture method.

Siegel and Fischl (1965) were concerned with pre-emergency training which prepares the public for a disaster or critical situation. They employed a technique known as "adjunct auto-instruction," which is meant to supplement other training techniques or points that need emphasis and stress. Adjunct auto-instruction tends to keep the learner active, and gives him feedback. The subjects were four matched groups ($N = 9$ to 13 per group) of semi-skilled, adult, employed women receiving attack survival material. The four experimental conditions provided that the subjects (*a*) receive material by phone, (*b*) read material in print, (*c*) read material in print and receive adjunct auto-instruction, or (*d*) receive material by telephone and receive adjunct auto-instruction. The non-adjunct groups were presented the material twice to equate for exposure time. A final examination administered at the end of training demonstrated that both adjunct types were significantly superior in promoting learning gains over non-adjunct materials ($p < .01$).

A CAI data management system was developed by Ford and Slough (1970) for an electronics course module. The course was tried out and revised three times using a total of 52 subjects. Next, the module was compared with normal classroom training using 51 CAI subjects and 200 traditional subjects. Afterwards, both groups took a standard school examination and a supplementary test. For all ability levels, CAI produced higher achievement than traditional classroom instruction. In addition, CAI produced time savings of 33 to 44 percent.

Showel, Taylor, and Hood (1966) constructed a leadership training package including tapes, filmstrips, and workbooks. This training package was used for an experimental group while a control group received traditional instruction (*i.e.*, lectures). The subjects were matched on the General Technical Aptitude area of the Army Classification Battery and randomly assigned to control and experimental groups. An essay examination was used to test achievement immediately after training and 10 weeks after training. The results demonstrated that the leadership automated package produced greater gain and was less costly than the conventional package.

Steadman, Bilinski, Coady, and Steinemann (1969) were interested in investigating alternate methods of training low-aptitude Naval personnel. Of 31 subjects, half were taught by instructor and half by programmed text. Achievement was measured by three quizzes and a practical performance test. Upon the termination of training, only eight subjects reached an adequate proficiency level in terms of the final practical performance test. These writers concluded that, in general, the course was not appropriate for low-aptitude personnel.

Programmer Characteristics

The selection of programmers for programmed learning is just as important as the selection of materials. Some of the characteristics of successful programmers are (*a*) "relatively high intelligence," (*b*) "interests in the area," (*c*) "attitudes favorable to the area and favorable to achieving the goal," (*d*) "compulsivity," and (*f*) "functional level of motivation (Melching, 1970, pp. 71-72)."

Television Instruction

TV instruction, although not used in the same way as CAI, is much less costly. TV instruction seems advantageous when instructor shortages exist, rapid dissemination of information is required, and student communication is not

necessary. This type of instruction is disadvantageous when applied lessons and student communication are needed.

Basic Education

Standlee and Hooprich (1962) feel that most tests of the effects of adult reading courses lack sophistication. Most experimenters measure reading ability before and after training, but fail to control for such factors as initial reading level, intelligence, motivation, equivalence of forms, test practice effects, set, test ceiling effects, change, regression effects, timed tests, type of test score, criterion choice, and differences between control and experimental subjects. These authors, after reviewing several sound studies, arrived at the following conclusions:

1. Reading speed gains are real. What happens to comprehension and vocabulary is uncertain, since they are confounded with speed. Eye movements usually improve.
2. Reading speed gains are retained. Generally, 60 to 70 percent was retained after six months to two years.
3. Reading instruction gains transfer to academic achievement, academic aptitude, clerical ability, and temperament. These gains may not be due to reading instruction, though, because these courses may also teach study skills, or give counselling and therapy which can also be associated with improvement.
4. No methods, materials, or programs of instruction were shown to be superior to any other. Also, no individual differences in personality, intelligence, or occupation were associated with reading skill gains.
5. Reading improvement courses are helpful for those whose jobs depend upon reading. In this case, increased speed is enough justification for taking the course.

Steinemann, Hooprich, Archibald, and Van Matre (1971) investigated the effects of a "wordsmanship" course given to 176 low-aptitude Naval personnel. These subjects characteristically have low verbal aptitude and unfavorable language attitudes which cause a bias against learning. Nevertheless, these investigators found that "the trainees substantially improved their knowledge and proficiency in each of the sub-course areas of wordsmanship, and most students reported a more favorable attitude toward words and a desire for self improvement of verbal skills."

Mollenkopf (1969) gave different 100-hour basic skills training courses (computation, spelling, filing, reasoning, paragraph meaning) to three different groups (office workers, laboratory technicians, and production employees). Most of the participants made sizable gains and most pretest and posttest score differences were significant, although regression and ceiling effects may have been involved. In almost all of the tests, at least 80 percent of the students made gains.

Hooprich and Steinemann (1966) indicated that there is "a general trend toward performance-oriented training courses in which technical mathematics and unnecessary electronics theory are minimized. . . . Increasing investigative attention devoted to performance evaluation problems is a reflection of the growing recognition of performance assessment as a critical factor in the final evaluation of total training effectiveness (pp. 17-18)."

Kent, Bishop, Byrnes, Frankel, and Herzog (1971a, 1971b) attempted to identify the Adult Basic Education (ABE) courses that were successful in job related settings (*e.g.*, obtaining job, promotions, entering training). Information was collected on 80 programs whose features or aspects were typed. Fifteen programs containing all features of interest were selected for the study. Checklist interviews were used to obtain data. The findings indicated that (*a*) there is a great need for ABE in basic abilities which vary from student to student and job market to job market; (*b*) the need for job related ABE is not being met in that the programs do not perform enough job placement, skill training, post instructional followup of students, self-evaluation, and improvement of materials; (*c*) theory, administration, and money are inadequate; (*d*) ABE programs should cooperate among themselves and with large centers for research; and (*e*) organizations should be invited to bid in order to conduct ABE job related programs.

Training Devices

Edgerton and Fryer (1950) have prepared a system for preliminary evaluation of a training aid. This system has the following features: (*a*) it is uniform and consistent; (*b*) it is brief; (*c*) it needs no special skills to administer; (*d*) it improves validity of technical judgments; (*e*) it shows advantages and defects of the training aid; (*f*) it provides for an overall judgment; and (*g*) it yields information from which an experimental evaluation of the training aid can be constructed.

Richardson, Bellows, Henry & Co. (1962) developed three evaluation forms for new training devices. These forms were constructed from literature reviews, descriptions of Navy devices, descriptions of industrial devices, and evaluation reports. These questionnaires were validated using the nomination technique in which instructors and training officers nominated devices as "best" or "worst." The resultant validity and reliability of the three methods proved adequate enough for use.

Siegel and Federman (1969) used Guilford's (1967) structure-of-intellect (SI) model to help derive the most appropriate aids and devices for training the tactical coordinator in the P-3c aircraft. Guilford's model allows the description of the mental tasks an operator performs in terms of intellectual load. These descriptions are quantitatively derived, and the needed aids and devices can be based upon them. The operations in the SI model specify the type of aids or devices for training. The contents in the SI model tell the subject matter of the aids or devices. Finally the SI products tell what is to be learned. The authors conclude that this technique defines training requirements and closes ". . . the loop between job analysis and the aid/device derivation."

Instructor Evaluation

A. Harris (1969) has found ". . . differences among teachers far more important than differences between methods and materials in influencing the reading achievement of children (p. 204)." The main criterion of teacher effectiveness should be pupil gain on standardized tests. The correlations between teacher ratings and tests are not large enough to support the use of ratings.

Bittner (1968) recently executed an interesting analysis of student evaluations of instructors. Subjective comments were collected from students on oral communication factors. These statements were content analyzed by six speech teachers (interrater reliability = .73). Five categories were derived: (*a*) rate of speaking, (*b*) volume, tone, and pitch, (*c*) use of audio-vidual aids, (*d*) use of discussion, and (*e*) organization of lecture. The largest number of comments concerned organization of lecture, while volume, tone, and pitch had the smallest number of comments. The most negative comments concerned volume, tone, and pitch, and the most positive concerned use of audio-visual aids. Rate of speaking was also somewhat negatively appraised. In addition, more negative comments were associated with graduate teaching assistants than with any other category.

Veldman and Peck (1969) wished to determine the influence on pupil evaluations of student teachers. These authors felt that the most reliable description of teacher behavior comes from the students. The Pupil Observation Survey (POSR) consisted of 38 items grouped into 10 scales. POSR data were collected on 554 student teachers at the University of Texas. The data were then factor analyzed, yielding five factors: (*a*) friendly and cheerful, (*b*) knowledgeable and poised, (*c*) lively and interested, (*d*) firm control, and (*e*) non-directive. Analysis of covariance was used to determine if five characteristics (grade in student teaching, grade of class, subject area, socio-economic status, level of school, and sex of teacher) had any effects. The results demonstrated that (*a*) all factors increased with increased student teaching grade; (*b*) only friendly-cheerful and lively-interested were positively and inversely related to grade level of students; (*c*) all factors except knowledgeable-poised were related to subject matter area; (*d*) as social class decreased, lively-interested increased, firm control decreased, and non-directive increased; and (*e*) females were rated higher on friendly-cheerful than males.

Hiller, Fisher, and Kaess (1969) performed a computer investigation of the verbal characteristics of effective classrooom lecturing. Fifty-five 15-minute lectures producing 105,000 words were analyzed for verbal fluency, optimal information amount, knowledge structure cues, interest, and vagueness. The findings demonstrated that vagueness in the lecture was most important. Vagueness is defined as ". . . the state of mind of a performer who does not sufficiently command the facts or the understanding required for maximally effective communication (p. 670)."

Military Research

Electronics Technicians. Applied Psychological Services (1971) recently developed a quick course of passive sonar training for system technicians. First, the training requirements were developed, followed by a course which was balanced between practical work and lecture presentation. Sonar technicians were given the course in one week. After finishing the course, they each completed a 13-item questionnaire. The mean value on a four-point scale for all 13 questions was 3.4. High values were concerned with the amount the student learned in the course. The authors concluded that this project was extremely useful, since it demonstrated that quickly but systematically developed courses could be useful.

Bilinski, Saylor, and Standlee (1969) used an analysis of on-the-job feedback to help increase training effectiveness. Electronics technician graduates were examined in regard to their ability to maintain a radar system. First, a job analysis was performed; then a structured interview was constructed from the job analysis to obtain information from a fleet sample of electronics technicians. This procedure elucidated difficult maintenance and problem areas for feedback into the training school.

Steinemann, Coady, Harrigan, and Matlock (1968) wanted to evaluate the job capabilities and fleet utilization of 64 four-year obligor graduates of electronics technician phase A-1 training. Performance measures and objective ratings were collected. Most electronics technicians were found to be more or less adequate. However, training limitations made on-the-job training and initial supervision necessary for all but the most routine tasks. Troubleshooting was found to be the weakest area. It was recommended that four-year obligors be given more training, or only be allowed to assist in fleet maintenance tasks. Steadman and Harrigan (1971) obtained similar results with six-year obligor data systems technicians. They suggest deemphasis of irrelevant electronics theory in favor of more practical training.

Helicopter Training. The studies discussed in this section were reviewed in a previous chapter of this report. The emphasis then was on dependent measures; now it is on evaluation.

Greer, Smith, and Hatfield (1967) wished to control for checkpilot personal bias in rating rotary wing students. The resultant ratings reflected the checkpilot's own standards rather than the student's flying skill. The training program was analyzed into maneuver components. Proficiency scales and instrument observation were substituted for the checkpilot's own method. The Pilot Performance Description Record (PPDR) was constructed to reflect the most critical aspects of each maneuver. The PPDR was administered to 50 advanced and 50 intermediate students. The results demonstrated that (*a*) reliability of flight proficiency evaluation improved; (*b*) the PPDR recorded specific student deficiencies; (*c*) checkpilots who were trained in PPDR were more consistent in their evaluation than checkpilots who were only oriented in PPDR; and (*d*) checkpilot training is necessary in the use of the PPDR.

Another approach, used by Greer (1968), to compensate for the variations in checkpilot standards involves grouping checkpilots with similar standards. Checkpilots were asked to complete an 11-point rating form, and those who agreed at .90 or better were paired together. In their actual evaluation duties, they correlated .65. It seems as though the earlier approach (Greer *et al.*, 1967) is more fruitful, since their checkpilots became better, less biased observers of behavior, while in this latter study (Greer, 1968), the checkpilots' bias is still allowed to operate.

Duffy (1968) and his associates (Duffy & Anderson, 1968; Duffy & Jolley, 1968) produced an objective and detailed scoring record. Students were scored on checkrides during and after training to yield a class percentage error. This procedure allows for class comparisons, grade comparisons, and instructor comparisons. If particular errors are identified among the students of one instructor, the instructor is given additional instructor training. Finally, if one checkpilot is more strict than the others, he is given counsel to make his observations more conforming.

Officer Training. Glickman and Vallance (1967) wished to find those aspects of the OCS curriculum which were most and least relevant to the job requirements of ensigns on destroyers. One-thousand critical incidents were collected and classifed as to "taught" and "not taught." Checklists containing 100 of the resultant items were sent to 30 to 50 high-level officers. They were required to judge the length of time in service after which the new officer should be able to handle the incident. The sooner an ensign was expected to cope with an incident, the more important that it be learned in OCS. Human relations, personnel administration, and leadership skills were found to be more important in this context than technical skills.

Morsh (1969) administered an officer management inventory to 10,242 Air Force officers who ranged in rank from lieutenant through colonel. The management inventory consists of a listing of tasks and duties, and a listing of military education topics. The officers rated, on an eight-point scale, the extent to which each task is a part of their job, and the extent to which each educational topic is useful in their job. Forty-three managerial types were derived from this analysis, although there was much overlap across types. The extent of managerial responsibility was directly related to officer grade. Also identified were training needs in leadership, communication, creative and logical thinking, problem solving, officer ethics, discipline and morale, and military customs and security. Other training topics were found to be of little use.

Task Analytic Methods. Stewart (1970) used task analysis to evaluate training effectiveness. Military task data were collected and analyzed to determine the extent to which it is job oriented. Stewart found that, in terms of cost, overtraining was as significant a problem as undertraining.

Siegel and Schultz (1961) and Siegel, Schultz, and Federman (1961) designed a system of training evaluation using matrix concepts. Essentially, training is acceptable if the average trainee performs with proficiency on a highly important task. Training is poor if the average worker performs poorly on a very important task and is very proficient on a task of low importance. This technique can yield a training index, an overtraining index, and an undertraining index for the entire training program. In addition, this method points to deficiencies in the program which need emphasis and parts of the program which need deemphasis. Schultz and Siegel (1962a, 1962b) applied the technique to posttraining performance of four Naval ratings. The results demonstrated that none of the groups were undertrained, while two of the groups seemed overtrained.

Aircraft Recognition. Whitmore, Cox, and Friel (1968) performed a study concerned with ground to air recognition training. The original training program for this aspect of aircraft recognition was thought to be inadequate. First, ground to air recognition slides were selected (16 Soviet and American jet fighter/attack aircraft). The paired-comparison method was employed to train in the discrimination. Eight-second exposures were given during training while five-second exposures were selected for the test. The results demonstrated that (*a*) 16 sessions were needed to achieve a 95 percent average recognition level; (*b*) class average on degraded images was 61 percent; (*c*) degraded images correlated .82 with the training achievement tests, indicating that the skill learned during training was not specific to the training slides; and (*d*) trainees maintained approximately the same position in class from achievement test to achievement test.

Summary

This chapter began with a discussion of some generally recognized trends. The most important trend seemed to be increased recognition of the multidimensionality of criterion measures. Next, there was a discussion of training needs and deficiencies followed by a very critical discussion of trends in cross-cultural training. This was followed by a presentation of some studies concerned with achievement measures as predictors of later success. Then there were reviews of studies involving sensitivity training, programmed instruction, CAI instruction, basic education, training and evaluation, and instructor evaluation. The final portion of this chapter was devoted to recent military research including electronics technician training, helicopter training, officer training, task analytic methods of evaluation, and aircraft recognition.

VI. COMPARATIVE EVALUATION

This chapter is divided into two parts. The first section involves comparative evaluation studies of non-low-aptitude men, while the second section focuses on low-aptitude evaluations. Generally, the studies reported here involve a relative comparison between two or more methods of instruction or training. In many cases, a new training method is compared with a standard method to determine if the latter should be replaced by the former.

Comparative Studies of Subjects Within Average or Higher Aptitude Ranges

Steinemann, Coady, Harrigan, Matlock, and Steadman (1969) compared six-year obligor electronics technicians with four-year obligors who are given less training. Six-year obligors were found to perform better on troubleshooting tests, test equipment examinations, written theory, and equipment tests. Questionnaire data on school limitations in troubleshooting were verified by the relative weakness found in this area as indicated by performance tests.

Hurlock (1971) grouped electronics technician training objectives into four short CAI lessons. Fifty randomly selected students were given CAI, and 180 were given traditional training. All subjects took the same final examination. The results demonstrated that overall achievement was 10 percent higher for CAI students. In addition, CAI instruction reduced training time 48.5 percent (17 hours to 8 3/4 hours).

Askren and Valentine (1970) were interested in the differences between Air Force instructors with job experience and without job experience in teaching a specialty area. The criteria used were student grades, student critiques, and supervisory evaluation. Seventy instructors and 585 students were used as subjects. Their conclusions were that (*a*) there were no significant differences in overall course grades across instructor type in a pneudraulics course; (*b*) there was an interaction for an environmental system course such that

grades of students from field experienced teachers increased from the beginning to the end of the course and decreased for non-field-experienced teachers from the beginning to the end of the course; (*c*) there were no significant differences in the student critiques; (*d*) field-experienced teachers were given an average supervisory rating of 3.22 (on a five-point scale) while non-field-experienced instruction received an average rating of 3.06; (*e*) a small number of the rating categories–knowledge of subject, student interest, and student participation–caused most of the difference; and (*f*) the job-experienced instructors were better at teaching theory. These investigators concluded that there is little practical difference in instructor type, but, if a shortage of field-experienced instructors exists, field-experienced persons should be used in practical, shop related courses.

Tallmadge (1968) attempted to study the interactions between trainee characteristics (*e.g.*, aptitudes and interests) and training methods. A one-week segment of Navy radarman school students was used as a setting for this experiment. In addition, a 32-item criterion test was developed. Three experimental conditions were involved: (*a*) subjects taught using rote memorization methods; (*b*) subjects taught problem solving, principles, and rationale approach; and (*c*) a standard approach, which is a mixture of other two methods. The 16 aptitude and interest measures did not interact with the three training methods as hypothesized. Perhaps the wrong training methods or the wrong aptitude and interest measures were used. It is also possible that other interactions existed which obscured the hypothesized interactions. Subjects in the rationale and understanding condition performed significantly better on the criterion test than the others, thus supporting the contention that this approach results in a hierarchically higher type of learning with better retention.

McFann, Buchanan, Lyons, Ward, and Waits (1958) compared a conventional Known Distance marksmanship training course with a new Trainfire I rifle marksmanship course. After four weeks of training, both groups received target detection and the Trainfire I marksmanship proficiency tests, as well as the conventional Known Distance test. The results demonstrated that Trainfire I training produced (*a*) a greater number of detected targets; (*b*) a shorter latency of target detection; (*c*) more target hits; (*d*) a higher percentage of men qualifying (the sum of marksman, sharpshooter, or expert); and (*e*) fewer qualifying as expert on the Known Distance range.

Olmstead (1968) compared Quick Kill Basic Rifle Marksmanship training (QKBRM) with traditional Basic Rifle Marksmanship training (BRM). QKBRM involves training the student to engage a target without aligning the sights of the weapon. Two experimental groups received QKBRM in their training and one control group received traditional BRM training (total $N = 824$). One of the experimental groups received a pre-training and a post-training questionnaire, and the other experimental group received only a post-training questionnaire. Control and experimental groups were compared on gains in confidence, attitude toward BRM, and drill sergeant attitudes toward QKBRM. Findings indicated an increase in confidence in both groups with QKBRM trainees gaining more confidence than traditional BRM trainees. The drill sergeant's attitude, though, was only somewhat favorable. One undeniable methodological weakness in this study is that the authors did not report any proficiency or marksmanship data across experimental groups.

Another study in this group concerns the effectiveness of an apparatus used as a simulator in driver training. The simulator-trained group was found to be superior in this experiment to the group trained on a projection-type driver trainer (Jeantheau & Anderson, 1966).

Caro and Isley (1966) used four groups of 33 subjects each in a study of Naval helicopter flight training. Groups A and B flew a training device 3.17 and 7.13 hours, respectively. Two control groups, C and C', received no device training. The Fisher exact probability test demonstrated that both device groups had fewer eliminations from training than did both control groups (10 percent to 30 percent at $p<.006$). In addition, the control groups had more unsatisfactory and below-average grades than did the two experimental groups.

In another study, Isley, Caro, and Jolley (1968) examined the advantage of a modified fixed wing device as a synthetic trainer for rotary wing procedures and aircraft control. Three groups of trainees were used each with 0, 10, and 20 hours, respectively, of synthetic training time. The experimenters found no difference in time to complete the course or in helicopter flight performance.

Isley (1968) and Isley and Caro (1969), in similar studies, examined the effects of a fixed wing rotary aircraft instrument trainer. Warrant officer candidates were divided into three treatments with 0, 10, and 20 hours, respectively, of synthetic training. The criteria used were deviations from regulation on 10 flight parameters in a

checkride. The results dramatically favored the group with no synthetic training in that they performed as well or better than the 20-hour group. The authors of this study seriously questioned use of the simulator.

Rhodes (1950) attempted to compare a new and an old ejection-seat trainer. The new trainer was more mobile, not as high, and more realistic in that it had a dummy cockpit. Training consisted of film, a lecture, and an ejection. Attitude was measured in both an "old" and a "new" group before and after ejection on each device. A group of reserve pilots was used as a control. No differences were found across groups; therefore, each is regarded as equally effective. Attitude did improve for both groups combined with reference to gain scores ($p<.01$). The author concluded that, regardless of device, overall ejection-seat training tends to increase confidence and decrease fear of this bailout method.

Gabriel and Burrows (1968) performed a study of pilot time-sharing performance. Time-sharing is concerned with alternating attention between two or more sources of information. Specifically, the pilot uses his instrument panel so much that he has little time to devote to outside scanning of the environment. The training task in this study was to improve the perception of midair threats of collision. The results suggested that use of the simulator can increase efficiency of pilot time-sharing between intra- and extra-cockpit stimuli.

Ward, Fooks, Kern, and McDonald (1970) wished to determine if the Basic Combat Training (BCT) and the Advanced Infantry Training (AIT) courses could be integrated for a sample of conscientious objectors in medical corpsman training. The content of the training courses currently used was catalogued. A job activities questionnaire was developed reflecting emergency medical care and secondary and recuperative treatment. The four types of tasks included in the training were company aidman, evacuation medic, aid-station dispensary medic, and ward nursing care medic. The criteria for selecting these groupings were availablity of supervision, frequency, and opportunity for on-the-job training. In the resultant 16-week course, practical work was emphasized and lecture was deemphasized. A large amount of TV instruction was used for 80 experimental students. For 80 other students, traditional training was involved. Combat proficiency, aidman proficiency, and attitude questionnaires were administered to all the trainees. In addition, an evaluation questionnaire was given to the instructors. The results of this effort demonstrated that (*a*) on military proficiency tests, both experimental and control groups performed equally well; (*b*) control subjects performed better on the Basic Combat Proficiency Test; (*c*) experimental subjects did better on physical skills used by medical corpsmen; (*d*) there were no significant differences in written knowledge tests; (*e*) experimental subjects performed better on medical performance tests; (*f*) experimental subjects had a higher opinion of the Army and its training than did standard subjects; and (*g*) instructors thought the experimental program was superior.

Judisch, Cooper, Francis, and Ray (1968) investigated the present curricula and job requirements of graduating medical corpsmen from two schools. They found that on knowledge tests San Diego students performed better on anatomy, physiology, first aid, and nuclear biological and chemical warfare. On the other hand, Great Lakes students were superior in patient care. A performance decrement was found over time such that, 24 weeks post-training, graduates were 10 percent worse than current students, and graduates of over 24 weeks were 16 percent worse. Also, a survey was performed to determine how much and where prior knowledge and information were acquired. Students reported gaining prior knowledge from lectures, films, readings, practical experience, and other visual aids. In all, though, this knowledge accounted for only 10 percent of the school knowledge. It was also found that San Diego students learned more from lectures than did Great Lakes students, and that Great Lakes students learned more from reality than did San Diego students. As a consequence of these results, the authors recommended revision in the curriculum.

Richlin, Federman, and Siegel (1958) compared general Naval technical training with a more specialized type of training under the Selective Emergency Service Rate Program (SESR). Each Naval rating in this program is subdivided and given a more specialized, shorter type of training. After training the men are utilized mostly in tasks for which they were trained. A Technical Behavior Checklist (TBCL) was developed as a criterion of performance for aviation machinist mates in the SESR program. Items for the TBCL were derived from tasks selected for their importance to the job, time consumed, and variability. The results of this study demonstrated that graduates of the

SESR program were equal to or better than the graduates of the more generalized program. Several other SESR studies were performed. In these studies it was demonstrated that (*a*) SESR trained air controllers performed as well as generally trained air controllers except in tower operations (Siegel, Richlin, & Federman, 1958); (*b*) SESR trained parachute riggers performed as well or better than generally trained parachute riggers (Siegel, Richlin, & Federman, 1958); and (*c*) SESR trained avionics technicians performed as well or better than pre-SESR trained avionics technicians (Richlin, Siegel, & Schultz, 1960),

Siegel, Federman, and Richlin (1959) administered a series of interviews to officers and petty officers in order to assess their opinion of the SESR program. One problem identified was the difficulty of assigning tasks to a more specialized man. Some supervisors felt SESR trained graduates achieved competence earlier, but that the more generally trained men were more useful.

CAI and TV Instruction. Gallagher (1970) attempted to investigate relevant learner characteristics and optimal types of instruction. He used four treatments: (*a*) computer assigned sequence of instruction–instructor evaluated product; (*b*) computer assigned sequence of instruction–computer evaluated product; (*c*) student selected sequence–instructor evaluated product; and (*d*) student selected sequence–computer evaluated product. Separate analyses of variance were conducted on the emergent data for four dependent variables: midterm examination, final product score, terminal or system time use, and time to complete cognitive portion of task. The results indicated that (*a*) there were no significant effects on any of the dependent measures; (*b*) both self-sequenced groups achieved superior performance on three of four dependent measures; (*c*) the computer assigned sequence of instruction was best in terms of cost; (*d*) those who performed best on the dependent measures were enthusiastic about the computer presentation; and (*e*) individual differences were minimized in the computer evaluated group. In conclusion, specific learner characteristics were related to success, and the student selected–computer evaluated approach was best in terms of costs.

Fishman, Keller, and Atkinson (1968) used CAI to present spelling drills to 29 fifth-grade students. Some words were presented via distributed practice, and other words were presented with massed practice. The results demonstrated that at the end of training the massed trials produced more correct responses, but 10 and 20 days later, the distributed practice group was superior ($p<.025$).

In another study, Rawls and Rawls (1968) found no significant differences in achievement and retention between conventional lecture presentation and closed circuit TV. College students, though, regarded the TV instruction unfavorably and preferred classroom instruction. This was true even among those who achieved high grades or had previous TV courses. The students were observed looking at the TV set only 20 percent of the time, while they looked at the lecturer 42 percent of the time.

Fidelity. Grimsley (1969a, 1969b) proposed to study the effects of variations in fidelity upon acquisition, transfer, and retention in group training procedures. There were 12 trainees per condition, trained in groups of four on the Nike-Hercules missile. They used a real (electric), a cold (non-electric), or an artist's sketch of the control panel. The subjects were tested immediately after training, four weeks later, and six weeks later on the 92-step missile firing procedure. No differences were found in training time, post-training performance, performance after four and six weeks, and in retraining time (after six weeks). This study suggests that a considerable saving of costs can be achieved by using a low-fidelity device. Similar results were found by Grimsley (1969a, 1969b) in a study that was identical except that group training procedures were not used.

Reduced Training Time. Longo and Mayo (1967) wished to determine if the 19-week airborne electronics training course could be decreased in time to 14 weeks. Two matched samples of trainees were used (total $N = 308$). The results proved disappointing since students in the longer course performed better than students in the shorter course.

Johnson and Salop (1968) observed that regular track avionics fundamentals training requires 16 weeks while accelerated track training needs only 10 weeks. The accelerated course differs from the standard course only in speed and amount of redundancy. In addition, only students of high ability are assigned to the accelerated track. It was found after training that accelerated students scored 2.6 points below students of the same ability on the single track program, but 5.9 points higher than all one track students, and 20.8 points higher than that required to graduate. The authors estimated that use of accelerated training in avionics fundamentals can save $750,000 a year.

Valverde (1969) decided to apply a systems approach to electronics maintenance training. First, behavior descriptions were derived from task analysis of the job requirements followed by the construction of performance tests based on the

objectives. Then a 14-week experimental training course was constructed for subjects with electronics aptitude scores ranging from the 60th to the 95th percentile. This group received only enough electronics theory to do the job. Another group with aptitude scores of 80 or better received the traditional 24-week course including 10 weeks of electronics principles. The experimental group was divided into two groups: 60th to 75th percentile and 80th to 95th percentile. The results demonstrated that (*a*) the high-aptitude experimental group performed better on the performance test than the medium-aptitude experimental group, which performed better than the traditionally trained control group; (*b*) the control group scored better on special theory and job knowledge tests; and (*c*) the cost of the experimental program was less than the cost of the traditional program.

Mental Health. Kumpan (1965) was interested in the effect of training on psychiatric aids in a mental hospital. The trainees consisted of 48 experimental subjects taking a four-month training program and 48 control subjects. There were two experimental wards of 30 patients each with the 48 experimental aids rotating among them. Kumpan found that the patients in the experimental wards did, indeed, improve. Psychiatric aids usually have the most contact with patients, but they are ill-qualified to help them because they do not understand the causes of mental illness.

Cochran and Steiner (1966) used an experimental group of 58 attendants for the retarded. They were given the Southern Regional Education Board Test before and after training. Sixteen control attendants were also used to determine if testing itself can cause a gain in posttest scores without training. Indeed, the control subjects gained 5.18 points ($p<.01$), while the experimental subjects gained 26.8 points ($p<.001$). Also, younger subjects with the least tenure seemed to make the greatest gains.

Poser (1966) performed an experiment to answer the question of whether special academic or intellectual knowledge is required to perform group therapy with schizophrenics. The three experimental conditions involved (*a*) 45 patients treated by psychiatrists and trained social workers, (*b*) 87 patients treated by students without any training, and (*c*) 63 untreated controls. All patients, before and after therapy, were given several tests to differentiate psychotic from normal, including tapping speed, reaction time, digit symbol, color-work conflict, verbal fluency, and the Verdun Association List. Analysis of covariance was performed on the data. The results indicated that (*a*) four of six tests showed significant gains by the lay therapist group as compared with the untreated groups; (*b*) two of six tests showed significant gains as the result of therapy by the professional therapist; and (*c*) three of six tests showed significant gains by the lay therapists over the professional therapists.

The conclusion from this experiment would seem to be that the use of lay therapists produced greater improvement than the professionally trained therapist. Of course, this involved only group therapy and not the traditional one-to-one situation in which a professional is most certainly needed.

Leadership Training. Rittenhouse (1953) compared two samples of enlisted men, one of which attended noncommissioned officer (NCO) leadership school. Both groups were compared on rank, assignment, and awards. The school group seemed to have a higher final rank and the non-school group had a greater gain in rank, but these differences were not statistically significant. The school graduate group had more infantry assignments (47.2 percent and 36.7 percent). Also, a greater proportion of the school graduate group received combat infantry badges.

Hood, Showel, and Stewart (1967) contrasted three methods of NCO leadership training with a non-training group. The trained leaders demonstrated (*a*) higher evaluations, (*b*) greater esprit de corps among their subordinates, (*c*) better proficiency test performance, (*d*) better preparation, briefing, and control of their men, and (*e*) more frequent structuring and use of rewards and definitions.

Barrett (1965) attempted to measure the impact of a 90-hour executive training program of the City of New York through comparison with a control group which did not undergo training (total $N = 255$). The results demonstrated no differences across groups in before- and after-performance ratings by peers and supervisors. The only measurable changes were increases in consideration and in initiating structure in the trainees and a decreased critical attitude toward subordinates.

Armor Training. The Human Resources Research Organization (Baker, Cook, Warnick, & Robinson, 1964) developed and evaluated a system for conducting tactical training of tank platoon crews. The tank crews themselves were trained on a miniature battlefield with radio

controlled tanks and simulated terrain. The tank commanders were trained on the Army Combat Decisions Game using tank models on a terrain board. A field performance test was then administered to the experimentally trained crews and to a group of matched controls. The crew receiving experimental training obtained significantly higher scores than the matched control crews.

Olson and Baerman (1955) wished to determine if a brief course in gas conservation had any effects on fuel consumption in the M48 tank. The three experimental conditions were (*a*) control–rotated among tanks in unit, (*b*) control–kept own tank, and (*c*) experimental–received instruction in fuel economy. These researchers found that the experimental group used less fuel when considerable stop-and-go driving was involved.

Reading and Verbal Instruction. Seventy-two scientists and engineers were trained for reading using a book method, and 42 were trained using mechanical machines (Jones & Carran, 1965). Different forms of the Diagnostic Reading Test were given before and after training. All subjects were found to have gained significantly after training, but in a followup 18 months later, the book approach was shown to be superior. In fact, performance of the machine trained group actually decreased after the time period, while performance of the book trained group continued to increase ($p<.002$).

Kelley and Mech (1967) wished to ascertain if a reading laboratory course could produce an increase in grade point average among college students. Twenty-three experimental subjects were matched with 23 controls. After three semesters no significant differences in grade point average were found. The investigators then divided their experimental and control groups by academic major. They found that (*a*) among education majors there was a statistically significant difference after three semesters ($p<.025$); (*b*) there was also a statistically significant difference among science and mathematics students ($p<.01$); and (*c*) there were no significant differences among social studies and literature majors. Perhaps, the education, science, and mathematics majors had an initially greater decrement in verbal ability, leaving a great deal more room for improvement. Also, education majors may have had a greater interest in reading improvement.

Frase (1969) taught 48 undergraduates verbal materials using two different methods of presentation. One method used a horizontal display of associations while the other used a vertical tabular display of associations. The results showed that the horizontal methods yielded superior learning, yet the subjects preferred the vertical tabular display.

Comparative Studies of Low-Aptitude Subjects

Skill Acquisition. Van Matre and Steineman (1966) trained 26 low-aptitude men in an electronics technician course in a shorter period of time and gave them skills more immediately useful on the job. This group was compared with 24 conventionally trained personnel in a fleet followup using performance tests, ratings, interviews, and written tests. The results demonstrated that the performance of the experimental group was adequate and not significantly different from the conventional group in proficiency.

Van Matre and Harrigan (1970) compared the performance of 54 marginally qualified electrical technicians with 51 well-qualified electrical technicians who underwent training. These groups were compared after they were on the job in the fleet for 24 months. A rating scale and a structured interview score were used as criteria. The conventionally trained men were rated as more capable in troubleshooting and use of test equipment, but were not generally rated differently from low-aptitude men. In fact, the lowest ratings obtained by low-aptitude men were average.

Mayo (1969) administered an aviation structural mechanic course to 30 Category IV personnel, i.e., the lowest 30 percent on the Armed Forces Qualification Test (AFQT). The fleet performance of this group was then compared with that of personnel who scored above the 30th percentile. Among the low-aptitude men, performance varied from highly satisfactory to unsatisfactory with no way of predicting which men would perform adequately. Low-aptitude men were found to have lower ratings ($p<.05$) than the other groups. Based on these results, Mayo suggested that Category IV personnel should not be used for this Naval rating unless there is a manpower shortage. It is noted, however, that the comparison group was given 25 percent more training and that ratings were used as criteria rather than performance tests.

Hooprich (1968) wished to determine the appropriateness of commissaryman training for Category IV personnel. The results, based on two studies, demonstrated that (*a*) 31 of 35 Category IV subjects successfully completed training, regardless of their low reading ability, although

their grades were significantly lower than the comparison group; (*b*) Category IV subjects needed to devote more outside time to study, and they required more time from instructors to meet criterion; (*c*) the differences across groups were most evident on paper-and-pencil tests and least evident on actual performance tests; (*d*) AFQT scores failed to predict school performance; and (*e*) reading test scores were significantly correlated with some aspects of performance.

Standlee and Saylor (1969) performed an equipment operator training study with Category IV subjects. The performance of six Category IV subjects was compared with 16 subjects who were not so classified. Then, the AFQT scores for this group and for commissaryman training were combined to determine if AFQT score predicted performance. It was found that (*a*) all Category IV subjects passed the course; (*b*) scores of the Category IV subjects were lower, especially on written tests as opposed to the more practical performance tests; (*c*) AFQT scores were unrelated to achievement; (*d*) mathematics was a source of trouble for Category IV personnel; and (*e*) Category IV men needed more individual attention and counselling.

Fox, Taylor, and Caylor (1969) compared the performance of low-aptitude men with higher aptitude men on several training tasks: visual monitoring, rifle assembly, missile preparation, phonetic alphabet, map plotting, and combat plotting. Low-aptitude groups needed 2 to 4 times as much training time, 2 to 5 times more training trials, and 2 to 6 times as much prompting to reach criterion. Middle-aptitude group performance was found to be more like that of the high-aptitude group than the low-aptitude group. The authors concluded that individual differences in aptitude must be recognized, and training programs must be designed to account for these differences.

Grunzke, Guinn, and Stauffer (1970) evaluated the performance of 26,915 low-aptitude men who were taken into the Air Force even though they were below the minimum acceptable level. The findings demonstrated that the low-aptitude men, as compared with subjects with higher aptitude, had (*a*) a smaller percentage completing basic training, (*b*) more disciplinary problems, (*c*) more unsuitable discharges, and (*d*) a lower percentage attaining skill level. In addition, among low-aptitude men, high school graduates and whites performed better than high school non-graduates and Negroes.

In another study, a manpower training program was surveyed by comparing 1,062 program graduates with 444 program dropouts (Trooboff, 1968). The results showed that 84 percent of the graduates received employment while only 67 percent of the dropouts received employment. Also, the average earnings of graduates increased from $.98 to $1.76 (79 percent), while the average earnings of dropouts increased from $1.07 to $1.51 (29 percent). Even though several factors were left uncontrolled, the author concluded that the program was successful.

Individualized Training. McFann (1969a, 1969b) found that the differences between high- and low-aptitude men in basic combat training were greatest on cognitive tasks and that the difference was not as marked on motor skills and proficiency tests, with most low-aptitude men meeting standard. In the study, high-, middle-, and low-aptitude groups were selected and trained, using videotape, a one-to-one student to teacher ratio, feedback, reinforcement, and small increments. In some tasks, low-aptitude men reached standard, but took 2 to 4 times longer, and in other cases they failed to master the material at all. McFann also found that aptitude interacts with method of instruction. The high-aptitude group was found to learn equally well with lecture or individualized training, while the low-aptitude group learned well with individualized training, but not with lecture.

J. Taylor (1970) found that both high- and low-aptitude personnel learn faster when given wire splice training via audiotape and slides as compared with a programmed book. For the high-aptitude personnel, the programmed book required 25 percent more training time; for the low-aptitude group, it took 50 percent more training time. From these results, Taylor suggests that training be adapted to individual differences.

Language Skills. Vineberg, Sticht, Taylor, and Caylor (1970) found that military training manuals were 6 to 8 grade levels above the reading level of Category IV personnel, and 4 to 6 grade levels above the reading level of higher aptitude subjects. Many of these individuals relied more heavily on asking and listening to others. In another study, Sticht (1969) found that among low-aptitude men learning by listening was more effective than learning by reading, although some did better by reading.

Summary

This chapter contained reviews of several comparative evaluation studies. Some of the studies were concerned with comparative evaluation of new training methods while others were concerned with methods of training low-aptitude personnel.

With regard to the training of low-aptitude men, more practical and individualized and less theoretical training seems superior to standard training procedures.

VII. DISCUSSION

There has been an increasing trend in the past decade in the use of factor analysis and other multivariate statistical techniques. Employment of these techniques has been made more feasible by the increased availability of high-speed computers. Many investigators, though, tend to use factor analysis as an end product or explanation rather than as an aid in data analysis. Factor analytic research can be misleading since the factors derived from the matrix reduction are directly dependent upon the variables making up the correlation matrix. This is a question of content validity. If the variable input is biased, then the results (factors) will be biased. In addition, most of the recent factor analytic literature has been so abstruse that it is difficult to understand the ideas presented, much less to implement them.

There has not been enough attention to canonical correlation, Q-factor analysis, and multivariate research design. No evaluative studies were found in which the first two of these methods were used, and too few studies using the latter were observed. Perhaps some of these sophisticated techniques are not appropriate to the data collected. In fact, a large portion of the data collected are not worthy of any analysis.

A large portion of the authors of the research studies reported in this review are guilty of violating one or more of the following canons of statistical methodology: (*a*) use of too few subjects; (*b*) use of inappropriate statistical techniques; (*c*) failure to use control groups, or use of inadequate controls; (*d*) use of improper sampling procedures; and (*e*) use of inappropriate, contaminated, or unreliable criteria.

Other quantitative methods which are given much lip service, but which are little used in practice except by their authors, are (*a*) sequential testing, (*b*) criterion-referenced testing, (*c*) confidence testing, (*d*) part correlation, (*e*) magnitude estimation, and (*f*) application of theory of signal detection. It behooves other investigators to try these techniques. Such methods can increase the sensitivity and generalizability of research findings.

One method which others are beginning to use is Campbell and Fiske's (1959) technique for establishing convergent and discriminant validity. Convergent validity exists if there is a high correlation between tests purporting to measure the same thing; and discriminant validity exists when tests measuring different factors are independent. This technique should prove very useful in the future for psychometricians involved in test construction and validation.

Another innovation which will come more into vogue is cost-effectiveness, or cost-benefit, analysis. This criterion is useful, as for as any other ratio, only if there is an adequate data base for both the numerator and the denominator of the ratio. Thus, the technique demands more precise economics and performance evaluative data.

Although the moderator variable technique is properly a subtopic under statistical methods, its emphasis in the recent literature demanded that it be given treatment in a separate chapter of this review. A test or measure can be a moderator variable when its use differentially determines the predictability of another test or measure. Almost any test score may be a potential moderator variable as are race, sex, personality, and other background factors.

Cognitive style seems to differ across deprived and non-deprived groups and must be accounted for and taken into consideration in order that the potential of the human resources in our society can be maximized.

Several studies were surveyed which use race and aptitude as moderator variables. One important conclusion (Boehm, 1971) to be drawn from this research is that objective and performance oriented dependent measures are less likely to show differences across racial groups than the more subjective rating methods. Another conclusion (McFann, 1969a, 1969b) is that high-aptitude groups learn equally well with lecture or individualized training, while low-aptitude groups learn well with individualized training but not with lecture.

Individualized or programmed instruction is another major educational trend which has achieved prominence in the last five or ten years. Individualized or programmed instruction represents an amalgam of the principles of learning theory with the idiosyncracies of the individual. Programmed instruction can be sequential, allowing the individual to proceed in very small steps through a fixed instructional sequence, or branched. Branching allows the individual's progress to be governed by his own responses.

Sequential testing has been used in individualized instruction in order to ascertain rapidly the level of knowledge possessed by the student. Also, criterion-referenced tests, rather than norm-referenced tests, have been employed, since the student must be able to perform each unit of instruction at a certain level of proficiency before advancing to the next unit of instruction.

Computer assisted instruction (CAI) is the application of computers to programmed instruction. CAI can be especially practical when a large number of short tests must be given to the trainee, and when instructor-student interaction is not considered crucial to learning.

Another noted trend was an increased concern with cross-cultural training and evaluation. Here, the "cultural assimilator" (Fiedler, Mitchell, & Triandis, 1970; Worchel & Mitchell, 1970) seemed to possess some merit. In this method, critical incidents are obtained regarding circumstances in which the norms of behaviors across cultures are quite different. Questions are asked about the incident, and the multiple-choice answer format is employed. The responses of a target sample from the host culture are employed to provide the correct answer keying.

Similarly, emphasis on increasing basic skills generally and reading skill specifically has achieved import. Courses in reading instruction have produced gains in reading speed, retention of reading speed, and transfer. No single method of reading instruction seems to have demonstrated superiority to another.

A method developed by Greer, Smith, and Hatfield (1967) has to some degree eliminated rater bias in helicopter checkpilots. After a task analysis, proficiency tests and instrument observation were substituted for the checkpilot's own evaluation method. This technique was able to (*a*) increase the reliability of evaluation, (*b*) identify specific student deficiencies, and (*c*) increase checkpilot consistency.

Siegel and Schultz (1961) and Siegel, Schultz, and Federman (1961) constructed an evaluative technique using matrix concepts which was successfully applied to a military setting (Schultz & Siegel, 1962). These writers feel that training is good if the average trainee performs proficiently on important tasks. Training is poor if the average worker performs poorly on important tasks. This method identifies deficiencies in the training program which need emphasis and those parts of the training program which need deemphasis.

The comparative studies discussed in this review were concerned with relative comparisons between two or more methods of instruction or training. In most cases a new training method was compared with a standard method to determine if the latter should be modified or replaced. Some of the conclusions to be drawn from this research are presented.

1. CAI is superior to standard instruction for electronics technicians in terms of achievement and speed (Hurlock, 1971).
2. If personnel shortages exist, job experienced Air Force instructors may be used in practical shop related courses, and instructors who are not job experienced may be used in lecture courses (Askren & Valentine, 1970).
3. Some of the newer Army marksmanship training methods are superior to the older, standard methods (McFann, Buchanan, Lyons, Ward, & Waits, 1958; Olmstead, 1968).
4. The benefits of simulator training are variable and seem to be dependent on a multiplicity of factors.
5. CAI, in the overall, seems to be a cost-effective training technique.
6. Students indicate a preference for traditional lectures over TV instruction (Fishman, Keller, & Atkinson, 1968).
7. Variations in the fidelity of a trainer seem to produce no observable performance differences.
8. Accelerated training is successful for high-aptitude students in avionics fundamentals training (Johnson & Salop, 1968).
9. NCO leadership training resulted in improved leader behavior over a no-training group (Hood, Showel, & Stewart, 1967).
10. Fuel conservation training can reduce fuel consumption in drivers of the M48 tank (Olson & Baerman, 1955).
11. A programmed book reading instruction course produces greater long-term improvement than machine training (Jones & Carran, 1965).

There has also been considerable recent concern with low-aptitude individuals who, generally, can perform many skilled tasks adequately when given

proper training. They tend to be slower learners and retain knowledge best when taught by practical rather than highly verbal means.

Finally, systematic approaches to evaluation and course development are beginning to receive some emphasis. These attempt to account for almost all of the variables that can affect training and student behavior. Most systems begin with a job analysis in order to derive a list of behaviorally oriented job requirements from which training objectives can be formulated. Many writers advocate a pre-training appraisal of the entering students in order to direct them to the training method which is most suited to their needs and abilities. Criterion-referenced tests and other measures of student behavior are then constructed in order to reflect the training objectives. Finally, after training, the students and the training program are evaluated through various means.

REFERENCES

AF Manual 50-2. *Instructional system development.* Washington: Department of the Air Force, 31 December 1970.

AF Manual 50-9. *Principles and techniques of instruction.* Washington: Department of the Air Force, 3 April 1967.

Abrams, A., & Pickering, E. *An evaluation of two short Vietnamese language courses.* San Diego: Naval Personnel and Training Laboratory, 1970.

Alkin, M. Evaluating the cost-effectiveness of instructional programs. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970. Pp. 221-238.

Allison, R. *Learning parameters and human abilities.* Princeton, N.J.: Educational Testing Service and Princeton University, 1960.

Allison, S. *The pilot training study: A cost estimating model for undergraduate pilot training.* Santa Monica, California: Rand Corporation, 1969.

Angell, D., Shearer, J., & Berliner, D. *Study of performance evaluation techniques.* Port Washington, N.Y.: U.S. Naval Training Device Center, 1964.

Applied Psychological Services. *Development and evaluation of a quick course in passive sonar training for AN/SQQ-23 (PAIR) system technicians.* Wayne, Pa.: 1971.

Askren, W., & Valentine, R. *Value of job experience to teaching effectiveness of technical training instructors.* AFHRL-TR-70-8. Wright-Patterson AFB, Ohio: Air Force Human Resources Laboratory, 1970.

Atkinson, J. Motivational determinants of risk-taking behavior. In J. Atkinson & N. Feather (Eds.), *A theory of achievement motivation.* New York: Wiley, 1966. Pp. 11-30.

Atkinson, R. Computer-based instruction in initial reading. *Proceedings of the 1967 invitational conference on testing problems,* 1967, 55-66.

Baker, F. Use of computers in educational research. *Review of educational research,* 1963, **33**(6), 566-578.

Baker, R., Cook, J., Warnick, W., & Robinson, J. *Development and evaluation of systems for the conduct of tactical training at the tank platoon level.* HumRRO Technical Report No. 88. George Washington University, 1964.

Baldwin, R., & Johnson, K. *An experiment in basic airborne electronics training. Part V. Evaluation of the revised courses.* San Diego: Naval Personnel Research Activity, 1968.

Bale, R., Rickus, G., & Ambler, R. *Replacement air group performance as a criterion for Naval aviation training.* Report No. NAMRL-1126. Pensacola, Florida: Naval Aerospace Medical Research Laboratory, 1970.

Barnes, J., & Statham, F. *U.S. Army primary helicopter school training program performance norms.* Aberdeen Proving Ground, Maryland: Human Engineering Laboratories. 1970.

Barrett, R. Impact of the executive program on the participants. *Journal of Industrial Psychology,* 1965, **3**(1), 1-13.

Bass, B., Thiagarajan, K., & Ryterband, E. *On the assessment of training value of small group exercises for managers.* Rochester, N.Y.: University of Rochester, 1968.

Becker, S. The parable of the pill. *Administrative Science Quarterly,* 1970, **15,** 94-96.

Beecroft, R. *The effectiveness of different training methods.* HumRRO Staff Memorandum, 1955.

Berdie, R. The uses of evaluation in guidance. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press: 1969. Pp 51-80.

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. Harris (Ed.), *Problems in measuring change.* Madison: University of Wisconsin, 1963.

Bergman, B. *The prediction of salesmen performance.* Unpublished paper, The Atlantic-Richfield Company, 1970.

Bergman, B., & Kujawski, C. *How to rate your subordinates.* Unpublished paper, The Atlantic-Richfield Company, 1969.

Bergstrom, H. *Job performance of young workers in relation to school background: A pilot approach toward using the job environment in evaluating both general and vocational education.* Minneapolis, Minnesota: Education Research and Development Council, College of Education, University of Minnesota, Office of Manpower Automation and Training, U.S. Department of Labor, Contract No. 81-22-30, 1968.

Biel, W. Training programs and devices. In R. Gagne (Ed.), *Psychological principles in system development.* New York: Holt, Rinehart & Winston, 1962. Pp. 343-383.

Bilinski, C., Saylor, J., & Standlee, L. *Training feedback on the AN/SPS-40 radar system.* San Diego: Naval Personnel Research Activity, 1969.

Bittner, J. Student evaluation of instructors' communication effectiveness. *College Student Survey,* 1968, **2**(2), 38-40.

Bligh, H. Trends in the measurement of educational achievement. *Review of Educational Research,* 1965, **35**(1), 34-52.

Bloom, B. Some theoretical issues relating to educational evaluation. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 6-25.

Bloom, B. Toward a theory of testing which includes Measurement-Evaluation-Assessment. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970. Pp. 25-50.

Boehm, V. Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology,* 1971, in press.

Bond, N., & Rigney, J. *Measurement of training outcomes.* Technical Report No. 66. University of Southern California, 1970.

Briggs, G., & Naylor, J. Team versus individual training, training task fidelity, and task organization effects on transfer performance by three-man teams. *Journal of Applied Psychology,* 1965, **49**(6), 387-392.

Brislin, R. *The content and evaluation of cross-cultural training programs.* Report No. P-671. Arlington, Virginia: Institute for Defense Analysis, 1970.

Bushan, V., & Ginther, J. Discriminating between a good and a poor essay. *Behavioral Science,* 1968, **13**(5), 417-420.

Campbell, D., & Fiske, D. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, **56,** 81-105.

Campbell, D., & Stanley, J. *Experimental and quasi-experimental designs for research.* New York: Rand McNally, 1963.

Campbell, J. *Personnel training and development.* Minneapolis, Minnesota: Minnesota University, Department of Psychology, 1970.

Campbell, J. Personnel training and development. *Annual Review of Psychology, 1971,* **22,** 565-602.

Campbell, J., & Dunnette, M. Effectiveness of T-group experiences in managerial training and development. *Psychological Bulletin,* 1968, **70,** 73-104.

Campbell, J., Dunnette, M., Lawler, E., & Weick, K. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

Carkhuff, R. Critical variables in effective counselor training. *Journal of Counseling Psychology,* 1969, **16**(3) 238-245.

Caro, P. *Flight evaluation procedures and quality control of training.* HumRRO Technical Report No. 68-3, 1968.

Caro, P., & Isley, R. *Changes in flight trainee performance following synthetic helicopter flight training.* HumRRO Professional Paper No. 1-66, 1966.

Carpenter, M., & Rapp, M. *The analysis of effectiveness of programs in elementary and secondary education.* Santa Monica, California: Rand Corporation, 1969.

Carss, B. *Systems analysis in education – a statement.* Educational Product Report, 1969, **2**, 43-44.

Carver, R. The curvilinear relationship between knowledge and test performance: Final examination as the best indicant of learning. In K. Wientge & P. Du Bois (Eds.), *Criteria in learning research.* Technical Report No. 9. St. Louis, Missouri: Washington University, 1966.

Carver, R. A model for using the final examination as a measure of the amount learned in classroom learning. *Journal of Educational Measurement,* 1969, **6**, 59-68.

Carver, R. Special problems in measuring change with psychometric devices. *Evaluative research: Strategies and methods.* Pittsburgh, Pa.: American Institutes for Research, 1970. Pp. 48-66.

Clarke, F. Confidence ratings, second choice responses, and confusion matrices in intelligibility tests. In J. Swets (Ed.), *Signal detection and recognition by human observers: Contemporary readings.* New York: Wiley, 1964. Pp. 620-648.

Cleary, T., Linn, R., & Rock, D. Reproduction of total test scores through the use of sequentially programmed tests. *Journal of Educational Measurement, 1968,* **5**, 183-187.(a)

Cleary, T., Linn, R., & Rock, D. An exploratory study of programmed tests. *Educational and Psychological Measurement,* 1968, **28**, 345-360. (b)

Cochran, I., & Steiner, K. Evaluation of an in service training program using the SREB information test. *American Journal of Mental Deficiency,* 1966, **70**(6), 913-917.

Cohen, L. Comments on Professor Messick's paper. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York; Holt, Rinehart & Winston, 1970. Pp. 204-210.

Cook, R., Hahn, C., & Sheppard, D. *Seminar program for professional and supervisory health care personnel: Development and evaluation.* Silver Spring, Md.: American Institutes for Research, 1971.

Coombs, C., Milholland, J., & Womer, F. The assessment of partial knowledge. *Educational and Psychological Measurement,* 1956, **16**, 13-37.

Crawford, M. Concepts of training. In R. Gagne (Ed.), *Psychological principles in system development.* New York: Holt, Rinehart & Winston, 1962. Pp. 301-339.

Crawford, M. *Research and development in training and education.* HumRRO Professional Paper No. 18-67, 1967.

Crawford, M. *Research in Army training: Present and future.* HumRRO Professional Paper No. 10-69, 1969.

Cronbach, L. Course improvement through evaluation. *Teachers College Record,* 1963, **64**, 672-683.

Danzig, E., & Keenan, J., Jr. *The development of procedures for evaluating the fleet proficiency of graduates of the Naval Air Technical Training Schools.* Philadelphia: Institute for Research in Human Relations, 1956.

Della- Piana, G., & Berger, M. A technique for evaluating the efficiency of programmed instruction. *Training and Development Journal,* 1970, **24**, 40-41.

Denova, C. Is this any way to evaluate a training activity? You bet it is! *Personnel Journal,* 1968, **4**(7), 488-493.

De Pauli, J., & Parker, E. *The introduction of the General Sonar Maintenance Trainer into Navy training for an evaluation of its effectiveness.* Technical Report No. 68-C-0005-1. Naval Training Device Center, 1969.

Dielman, T., & Wilson, W. Convergent and discriminant validity of three measures of ability, aspiration-level, achievement, adjustment, and dominance. *Journal of Educational Measurement,* 1970, **7**, 185-190.

Du Bois, P. *Multivariate correlational analysis.* New York; Harper & Brothers, 1957.

Duffy, J. *A quality control program applied to helicopter training.* Collected papers prepared under work unit LIFT: Army Aviation Helicopter Training. HumRRO Professional Paper No. 18-68, 1968, 11-12.

Duffy, J., & Anderson, E. *Flight training quality control.* Collected papers prepared under work unit LIFT: Army Aviation Helicopter Training. HumRRO Professional Paper No. 18-68, 1968, 13-20.

Duffy, J., & Jolley, O. Briefing on task LIFT. Collected papers prepared under work unit LIFT: Army Aviation Helicopter Training HumRRO Professional Paper No. 18-68, 1968, 3-10.

Dunnette, M. A modified model for test validation and selection research. *Journal of Applied Psychology,* 1963, 47, 317-323.

Eddy, W., Glad, D., & Wilkins, D. Organizational effects on training: A study conducted on a public administration program. *Training and Development Journal,* 1967, 15-23.

Edgerton, H., & Fryer, D. *The development of an evaluation procedure for training aids and devices.* New York: Richardson, Bellows, Henry, & Co., 1950.

Edwards, C. The performance rating: A study in teacher evaluation. *Educational and Psychological Measurement,* 1968, **28**(2), 487-492.

Englemann, S. Relating operant techniques to programming and teaching. *Journal of School Psychology,* 1968, **6**(2), 89-96.

Federman, P., & Siegel, A. *Communication as a measureable index of team behavior.* Wayne, Pa.: Applied Psychological Services, 1965.

Ferguson, R. *Computer-assisted criterion-referenced testing.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1969.

Ferguson, R. *Computer-assisted criterion-referenced measurement.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1970.

Fiedler, F., Mitchell, T., & Triandis, H. *The culture assimilator: An approach to cross-cultural training.* Seattle: Washington University, 1970.

Fishman, E., Keller, L., & Atkinson, R. Masses versus distributed practice in computerized spelling drills. *Journal of Educational Psychology,* 1968, **59**(4), 290-296.

Fiske, D. Why do we use situational performance tests? *Personnel Psychology,* 1954, **7**, 464-469.

Fitzpatrick, R. The selection of measures for evaluating programs. *Evaluative research: Strategies and methods.* Pittsburgh: American Institutes for Research, 1970. Pp. 67-81.

Flanagan, J. The uses of educational evaluation in the development of programs, courses, instructional materials and equipment. Instructional and learning procedures, and equipment. Instructional and learning procedures and administrative arrangements. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 221-241.

Flaugher, R., Campbell, J., & Pike, L. *Ethnic group membership as a moderator of supervisor's ratings.* Princeton: Educational Testing Service, 1969.

Foley, P. *Validity of the OQT for minority group applicants to officer candidate school.* Naval Personnel Research and Development Laboratory. WRR 71-1, 1971 (Revised).

Ford, J., & Slough, D. *Development and evaluation of computer assisted instruction for Navy electronics training: 1. Alternating current fundamentals.* San Diego: Naval Personnel and Training Research Laboratory, 1970.

Forrest, F. *Develop an objective flight test for the certification of a private pilot.* Daytona Beach: Embry-Riddle Aeronautical Institutes, 1970.

Fostvedt, D. Criteria for the evaluation of high school English composition. *Journal of Educational Research,* 1965, **59**(3), 108-112.

Fox, W., Taylor, J., & Caylor, J. *Aptitude level and the acquisition of skills and knowledges in a variety of military training tasks.* Technical Report 69-6. The George Washington University, HumRRO Division No. 3, 1969.

Frase, L. Tabular and diagrammatic presentation of verbal materials. *Perceptual and Motor Skills,* 1969, **29**, 320-322.

French, J. *The relationship of problem solving styles to the factor composition of tests.* Princeton: Educational Testing Service, 1963.

Furno, O. Sample survey designs in education-focus on administrative utilization. *Review of Educational Research,* 1966, **36**, 552-565.

Gabriel, R., & Burrows, A. Improving time-sharing performance of pilots through training. *Human Factors,* 1968, **10**(1), 33-40.

Gagne, R. Human functions in systems. In R. Gagne (Ed.), *Psychological principles in system development.* New York: Holt, Rinehart & Winston, 1962. Pp 35-74.

Gagne, R. (Ed.) *Psychological principles in system development.* New York: Holt, Rinehart & Winston, 1962.

Gagne, R. Curriculum research and the promotion of learning. *Perspectives of curriculum evaluation.* Chicago: Rand McNally & Co., 1967. Pp. 19-38.

Gagne, R. Computer assisted instruction: Some facts and fancies, a discussion. In P. Dubois (Ed.), *Psychological research in adult learning.* St. Louis, Missouri: Washington University, 1968. Pp. 27-29.

Gagne, R. Instructional variables and learning outcomes. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970. Pp. 105-126.

Gallagher, P. *An investigation of instructional treatments and learner characteristics in a computer-managed instruction course.* Technical Report No. 12. Tallahassee, Florida: Florida State University CAI Center, 1970.

Gardner, W. *The use of confidence testing in the academic instructor course.* Lexington, Mass.: Shuford-Massengill Corporation, 1970.

Gay, L. *An investigation into the differential effectiveness for males and females of three CAI treatments on delay retention of mathematical concepts.* Technical Report No. 12. Tallahassee, Florida: Florida State University CAI Center, 1969.

Geisert, P. *A comparison of the effects of information mapped learning materials on the learning of concepts via the printed page and computer cathode ray tube.* Technical Memorandum No. 24. Tallahassee, Florida: Florida State University CAI Center, 1970.

Gerken, C. An objective method for evaluating training programs in counseling psychology. *Journal of Counseling Psychology,* 1969, **16**(3), 227-237.

Gideonse, H. The relative impact of instructional variables: The policy implications for research. *Record,* 1968, **69**(7), 625-640.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 1963, **18**, 519-521.

Glaser, R. Adapting the elementary school curriculum to individual performance. *Proceedings of the 1967 invitational conference on testing problems,* 1967, 3-36.

Glaser, R. *Individual differences in learning.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1970. (a)

Glaser, R. Evaluation of instruction and changing educational models. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York; Holt, Rinehart & Winston, 1970. Pp. 70-86. (b)

Glaser, R., & Cox, R. *Criterion-referenced testing for the measurement of educational outcomes.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1968.

Glaser, R., & Klaus, D. Proficiency measurement: Assessing human performance. In R. Gagne (Ed.), *Psychological principles in system development.* New York: Holt, Rinehart & Winston, 1962. Pp. 419-472.

Glaser, R., & Glanzer, M. *Abstract of training and training research.* Pittsburgh: American Institutes for Research, 1958.

Glaser, R., & Nitko, A. *Measurement in learning and instruction.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1971.

Glickman, A., & Vallance, T. Curriculum assessment with critical incidents. In E. Fleishman (Ed.), *Studies in personnel and industrial psychology.* Homewood, Ill.: The Dorsey Press, 1967. Pp. 189-198.

Golembiewski, R., & Carrigan, S. Planned change in organizational style based on the laboratory approach. *Administrative Science Quarterly,* 1970, **15**, 79-93.

Goolsby, T., Frary, R., & Lasco, R. Factorial structure and principal correlates of the Florida Bar Examination. *Educational and Psychological Measurement,* 1968, **28**(2), 427-432.

Gorth, W., & Grayson, A. A program to compose and print tests for instructional testing using item sampling. *Educational and Psychological Measurement,* 1969, **29**, 173-174.

Grant, O., & Bray, O. Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology,* 1970, **54**(1), 7-14.

Greer, G. *The effects of flight proficiency measurement reliability of differences on check pilot standards.* Collected papers prepared under work unit LIFT: Army Aviation Helicopter Training. HumRRO Professional Paper No. 18-68, 1968, 1-2.

Greer, G., Smith, W., & Hatfield, J. *Improving flight proficiency evaluation in army helicopter pilot training.* HumRRO Technical Report No. 77, 1967.

Grimsley, D. *Acquisition, retention, and retraining: Effects of high and low fidelity in training devices.* HumRRO Technical Report No. 69-1, 1969. (a)

Grimsley, D. *Acquisition, retention, and retraining: Group studies on using low fidelity training devices.* HumRRO Technical Report No. 69-4, 1969. (b)

Gronlund, N. *Constructing achievement tests.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Grunzke, M.E., Guinn, N., & Stauffer, G.F. *Comparative performance of low-ability airmen.* AFHRL-TR-70-4, AD-705 575. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, January 1970.

Gubins, S. *The impact of age and education on the effectiveness of training: A benefit-cost analysis.* Baltimore, Maryland: Johns Hopkins University, 1970.

Guilford, J. *The nature of human intelligence.* New York: McGraw-Hill, 1967.

Guilford, J. Comments on Professor Bloom's paper. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970. Pp. 62-66.

Guinn, N., Tupes, E. & Alley, W. *Demographic differences in aptitude test performance.* AFHRL-TR-70-15, AD-710 618. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, May 1970. (a)

Guinn, N., Tupes, E.C., & Alley, W.E. *Cultural subgroup differences in the relationship between Air Force aptitude composites and training criteria.* AFHRL-TR-70-35, AD-715 922. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, September 1970. (b)

Gutsch, K. Instrumental music performance: One approach toward evaluation. *Journal of Educational Research,* 1966, **59**(8), 377-380.

Haggard, D., & Willard, N. *An experimental program of instruction on the management of training.* HumRRO Technical Report No. TR-70-9, 1970.

Hammel, D. *Flexible teaching methods for CAI systems.* Austin, Texas: Texas University Electronics Research Center, 1969.

Hansen, D., Hedl, J., & O'Neal, H. *Review of automated testing.* Technical Memorandum No. 30. Tallahasee, Florida: Computer Assisted Instruction Center, 1971.

Harris, A. The effective teacher of reading. *Reading Teacher,* 1969, (**3**), 195-204.

Harris, C. Comments on Professor Wiley's paper. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970. Pp. 269-271.

Hawkridge, D. Designs for evaluative studies. *Evaluative research: Strategies and methods.* Pittsburgh, Pa.: American Institutes for Research, 1970. Pp. 24-47.

Hemphill, J. The relationship between research and evaluation studies. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 189-220.

Hess, R., & Shipman, V. Early experience and the socialization of cognitive modes in children. *Child Development,* 1965, **36**, 869-886.

Hiller, J., Fisher, G., & Kaess, W. A computer investigation of verbal characteristics of effective classroom lecturing. *American Educational Research Journal,* 1969, **6**, 661-675.

Holtzman, W. The changing world of mental measurement and its social significance. *American Psychologist,* 1971, **26**, 546-553.

Hood, P., Showel, M., & Steward, E. *Evaluation of three experimental systems for noncommissioned officer training.* HumRRO Technical Report No. 67-12, 1967.

Hooprich, E. *A second investigation of the feasibility of Navy commissary man training for group IV personnel.* San Diego, California: Naval Personnel Research Activity, 1968.

Hooprich, E., & Steinemann, J. *A review of electronics training research literature.* San Diego, California: Naval Personnel Research Activity, 1966.

Horn, J. Some characteristics of classroom examinations. *Journal of Educational Measurement*, 1966, **3**, 293-295.

House, R. Leadership training: Some dysfunctional consequences. *Administrative Science Quarterly*, 1968, **12**(4), 556-571.

Howard, A., & Correll, P. Evaluating psychological interns. *Journal of Counseling Psychology*, 1966, **30**(1), 78-80.

Hunter, M., Lyons, J., MacCaslin, E., Smith, R., & Wagner, H. *The process of developing and improving course content for military technical training.* Alexandria, Virginia: George Washington University, Human Resources Research Office, 1969.

Hurlock, R. *Development and evaluation of computer assisted instruction for Navy electronics training. 2. Inductance.* Research Report No. SRR 71-22. San Diego, California: Naval Personnel and Training Research Laboratory, 1971.

Husen, T. International impact of evaluation. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969. Pp 335-350.

Isley, R. *Inflight performance after zero, ten, or twenty hours of synthetic instrument flight training.* HumRRO Professional Paper No. 23-68, 1968.

Isley, R., & Caro, P. *Evaluation of synthetic instrument flight training in the officer/warrant officer rotary wing aviator course.* HumRRO Professional Paper, 1969.

Isley, R., Caro, P., & Jolley, G. *Evaluation of synthetic instrument flight training in the officer/warrant officer rotary wing aviator course.* HumRRO Technical Report No. 68-14, 1968.

Jaeger, R. School testing to test the schools. *Proceedings of the 1970 invitational conference on testing problems*, 1970, 39-52.

Jeantheau, G., & Anderson, B. *Training system use and effectiveness evaluation.* Port Washington, N.Y.: U.S. Naval Training Device Center, 1966.

Jenkins, J., Ewart, E., & Carroll, J. *The combat criterion in naval aviation.* Washington, D.C.: National Academy of Sciences, National Research Council, 1950.

Jensen, A. "How much can we boost I.Q. and scholastic achievement?" *Harvard Educational Review*, 1969, 39.

Jensen, A. Individual differences in visual and auditory memory. *Journal of Educational Psychology*, 1971, **62**(2), 123-131.

Johnson, G. The purpose of evaluation and the role of the evaluator. *Evaluative research: Strategies and methods.* Pittsburgh, Pa.: American Institutes for Research, 1970. Pp. 1-23.

Johnson, K. Evaluation of a self-paced course. In P. DuBois (Ed.), *Psychological research in adult learning.* St. Louis, Missouri: Washington University, 1968. Pp. 55-59.

Johnson, K. *Identification of difficult units in a training program.* Technical Bulletin STB 69-4. San Diego, California: Naval Training Research Laboratory, 1969. (a)

Johnson, K. *Retention of electronic fundamentals: Differences among topics.* San Diego, California: Naval Personnel Research Activity, 1969. (b)

Johnson, K., & Salop, P. *Two track training for avionics fundamentals.* San Diego, California: Naval Personnel Research Activity, 1968.

Jolley, O., & Caro, P. A determination of selected costs of flight and synthetic flight training. HumRRO Technical Report No. TR-70-6, 1970.

Jones, D., & Carran, T. Evaluation of a reading development program for scientists and engineers. *Personnel Psychology*, 1965, **18**(3), 281-295.

Judisch, J., Cooper, R., Francis, P., & Ray, T. *Evaluation of the basic hospital corps school.* State College, Pa.: HRB-Singer, Inc., 1968.

Kaplan, H., Freedman, A., & Kaplan, H. The evaluation of psychiatric residents by objective multiple-choice examinations. *American Journal of Psychiatry*, 1968, **124**(8), 1101-1106.

Katz, I. Some motivational determinants of racial differences in intellectual achievement. *International Journal of Psychology*, 1967, **2**(1), 1-12.

Kavanagh, M., MacKinney, A., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin,* 1971, **75,** 34-49.

Keller, F. "Good-bye, teacher. . . " *Journal of Applied Behavior Analysis,* 1968, **1**(1), 79-89.

Kelley, C., & Kelley, E. *A manual for adaptive techniques.* Santa Monica, California: Dunlop and Associates. AR-196-050, 1970.

Kelley, I., & Mech, D. The relationship between college reading laboratory experience and gains in college grade point average. *Journal of the Reading Specialist,* 1967, **7**(2), 50-54.

Kent, W., Bishop, R. Byrnes, M., Frankel, S., & Herzog, J. *Job related adult basic education, volume I. Summary and recommendations.* Report SDC-TM-WO-368-OEO-LA-860. Falls Church, Virginia: Systems Development Corp., 1971. (a)

Kent, W., Bishop, R., Byrnes, M., Frankel, S., & Herzog, J. *Job related adult basic education, volume II. Approach and detailed findings.* Report SDC-TM-WO-369-OEO-LA-861. Falls Church, Virginia: Systems Development Corp., 1971. (b)

Klein, S., & Evans, F. An examination of the validity of nine experimental tests for predicting success in law school. *Educational and Psychological Measurement,* 1968, **28**(3), 909-913.

Kumpan, H. How effective are aide training programs? *Mental Hospitals,* 1965, **16**(7), 209-211.

Lavisky, S. *HumRRO research and the Army's training programs.* HumRRO Professional Paper No. 36-69, 1969.

Likert, R. *The human organization.* New York: McGraw-Hill, 1967.

Lindvall, C., & Cox, R. The role of evaluation in programs for individualized instruction. In R. Tyler (Ed.), *Educational evaluations: New roles, new means.* Chicago: University of Chicago Press, 1969, 156-188.

Longo, A., & Mayo, D. *Effect of reduction in training time upon knowledge of electronics fundamentals.* Technical Bulletin STB 67-3. San Diego, California: U.S. Naval Research Activity, 1967.

Lord, F. *A theoretical study of two-stage testing.* Princeton, N.J.: Educational Testing Service, RB-69-95, 1969.

Lord, F. Robbins-Munro procedures for tailored testing. *Educational Psychological Measurement,* 1971, **31**(1), 3-31. (a)

Lord, F. *Tailored testing, an application of stochastic approximation.* Princeton, N.J.: Educational Testing Service, RM-71-2, 1971. (b)

Lortie, D. The cracked cake of educational custom and emerging issues in evaluation. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970. Pp. 149-164.

Lumsdaine, A. Comments on Professor Glasser's paper. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970. Pp. 91-95.

Lunneborg, P., & Lunneborg, C. Undergraduate preparation and success in law school. *Vocational Guidance Quarterly,* 1967, **15**(3), 196-200.

Mager, R. *Teaching: Today and tomorrow.* Collected papers prepared under work unit TEXTRUCK. Methods of instruction in technical training. HumRRO Professional Paper No. 34-70, 1970, 1-11. (a)

Mager, R. *Preliminary studies in automated teaching.* Collected papers prepared under work unit TEXTRUCK. Methods of instruction in technical training. HumRRO Professional Paper No. 34-70, 1970-12. (b)

Manuel, E. *Human resources management system. Project evaluation, model I.7.* Ozarks Regional Commission, Final Report, 1970.

Martin, H. The assessment of training. *Personnel Management,* 1957, **39**, 88-93.

Massengill, H. & Shuford, E. *Confidence testing at the academic instructor course of the Air University.* Lexington, Mass.: Shuford-Massengill, Corp., 1969.

Mayo, A. *Fleet performance of project 100,000 personnel in aviation structural mechanic S (structures) rating.* Research Report SRR-69-17, 1969.

McFann, H. *Progress report on HumRRO research on project 100,000.* HumRRO Professional Paper No. 25-69, 1969. (a)

McFann, H. *Individualization of Army training. Innovations for training.* HumRRO Professional Paper No. 6-69, 1969, 1-9. (b)

McFann, H., Buchanan, D., Lyons, J., Ward, J., & Waits, C. *Extension of research in trainfire I basic rifle marksmanship course.* Alexandria, Virginia: George Washington University, Human Resources Research Office, 1958.

McGuire, C. An evaluation model for professional education-medical evaluation. *Proceedings of the 1967 invitational conference on testing problems,* 1967, 37-52.

McGuire, C. Testing in professional education. *Review of Educational Research,* 1968, **38**(1), 49-60.

McGuire, C., & Babbott, D. Simulation technique in the measurement of problem solving skills. *Journal of Educational Measurement,* 1967, **4**, 1-9.

McSheehy, D. *Performance evaluation of apprentice fuel specialist graduates of air training command course No. ABR 64330A.* Report No. APGO-TR-59-11. Eglin AFB, Fla.: Air Proving Ground Center, 1959.

Melching, W. *Some research needs in selecting and training programmers.* Collected papers prepared under work unit TEXTRUCK. Methods of instruction in technical training. HumRRO Professional Paper No. 34-70, 1970, 69-73.

Merrifield, P. Trends in the measurement of special abilities. *Review of Educational Research,* 1965, **35**(1), 25-33.

Merwin, J. Historical review of changing concepts of evaluation. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 6-25.

Messick, S. The criterion problem in the evaluation of instruction: Assessing possible, not just intended, outcomes. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970. Pp. 183-202.

Miehle, W. & Siegel, A. *Development of performance evaluation measures. Personnel psychophysics: Quantification of malfunction detection probability.* Wayne, Pa.: Applied Psychological Services, 1967.

Miller, W. & Norris, R. Entrance age and school success. *Journal of School Psychology,* 1967, **6**, 47-60.

Molenkopf, W. Some results of three basic skills training programs in an industrial setting. *Journal of Applied Psychology,* 1969, **53** 343-347.

Morsh, J. *Survey of Air Force officer management activities and evaluation of professional military education requirements.* AFHRL-TR-69-38, AD-705 574. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, December 1969.

Naylor, J., Briggs, G., & Reed, W. Task coherence, training time, and retention interval effects on skill retention. *Journal of Applied Psychology,* 1968, **52**(5), 386-393.

O'Brien, G., Fiedler, F., & Hewett, T. *The effects of programmed culture training upon the performance of volunteer medical teams in Central America.* Urbana, Ill.: Illinois University Group Effectiveness Research Laboratory, 1969.

Olmstead, J. *The effects of "quick kill" upon training confidence and attitudes.* HumRRO Technical Report No. 68-15, 1968.

Olson, M., & Baerman, D. *The effect of fuel conservation training on M 48 tank gasoline consumption.* Alexandria, Virginia: George Washington University, Human Resources Research Office, 1955.

Osborn, W. *An approach to the development of synthetic performance tests for use in training evaluation.* Alexandria, Virginia: HumRRO Professional Paper No. 30-70, 1970.

Peck, R., & Dingman, H. Some criterion problems in evaluation of teacher education. *Psychological Reports,* 1968, **23**(1), 300.

Pfeiffer, M., & Siegel, A. *Post-training performance, criterion development and application. Personnel psychophysics: Initial studies into psychological scaling of electronics job complexity.* Wayne, Pa.: Applied Psychological Services, 1965.

Pfeiffer, M., & Siegel, A. *Post-training performance, criterion development and application: Personnel psychophysics: A model of the job of the Naval avionics personnel and further studies in personnel psychophysics.* Wayne, Pa.: Applied Psychological Services, 1966.

Pfeiffer, M., & Siegel, A. *Development of performance evaluative measures. Personnel psychophysics: The relationship between structure of intellect scale values and job complexity.* Wayne, Pa.: Applied Psychological Services, 1967. (a)

Pfeiffer, M., & Siegel, A. *Post-training performance, criterion development and application: Personnel psychophysics: The functional relationship between job complexity and number of electronic maintenance training variables.* Wayne, Pa.: Applied Psychological Services, 1967. (b)

Pollack, I., & Decker, L. Confidence ratings, message reception, and the receiver operating characteristic. In J. Swets (Ed.), *Signal detection and recognition by human observers: Contemporary readings.* New York: Wiley, 1964. Pp. 592-608.

Popham, W. & Husek, J. Implication of criterion-referenced measurement. *Journal of Education Measurement,* 1969, **6**, 1-9.

Poser, E. The effect of therapists' training on group therapeutic outcome. *Journal of Counsulting Psychology,* 1966, **30**(4), 283-289.

Project Impact–Computer administered instruction: Description of the hardware/software subsystem. HumRRO Technical Report No. 70-22, 1970.

Provus, M. Evaluation of ongoing problems in the public school system. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 242-283.

Rawls, J., & Rawls, D. Evaluation of closed-circuit television in teaching educational psychology. *Psychological Reports,* 1968, **22**(3), 1041-1044.

Rhea, B. *The relationship of OCS grades to officer fitness report marks.* San Diego, California: U.S. Naval Personnel Research Activity, 1965.

Rhodes, C. *Effectiveness of ejection seat training with special reference to SDC device No. 6EQ-2.* Richardson, Bellows, Henry & Co., Inc., SCD Human Engineering Project 20-A-6, 1950.

Richards, J., Holland, J., & Lutz, S. Prediction of student accomplishment in college. *Journal of Educational Psychology,* 1967, **58**(6), 343-355.

Richardson, Bellows, Henry, & Co. *Development of evaluation procedures for prototype devices.* New York: 1962.

Richlin, M., Federman, P., & Siegel, A. *Post-training performance, criterion development and application: Development and application of a TBCL criterion to the SESR program for jet aviation machinist's mates.* Wayne, Pa.: Applied Psychological Services, 1958.

Richlin, M., Siegel, A., & Schultz, D. *Post-training performance, criterion development and application: Development and application of a TBCL criterion to the SESR program for aviation electronics technicians.* Wayne, Pa.: Applied Psychological Services, 1960.

Rimland, B. *The search for measures of practical intelligence: Research project 100,000.* Paper presented at the American Psychological Association Meeting, Washington, D.C., 3 September 69.

Rittenhouse, C. *A follow-up study of NCO leaders school graduates.* Alexandria, Virginia: George Washington University, Human Resources Research Office, 1953.

Rotter, J. Generalized expectancies for internal vs. external control of reinforcement. *Psychological Monographs,* 1966, **80**(1), (Whole No. 609).

Rundquist, E. The prediction ceiling. *Personnel Psychology,* 1969, **22**, 109-116.

Ryan, J. Previous instructional program as a moderator of the predictive validity of college entrance tests in mathematics. *Educational and Psychological Measurement,* 1968, **28**(3), 937-941.

Schultz, D., & Siegel, A. *Post-training performance criterion development and application: A selective review of methods for measuring individual differences in on-the-job performance.* Wayne, Pa.: Applied Psychological Services, 1961. (a)

Schultz, D., & Siegel, A. Generalized Thurstone and Guttman scales for measuring technical skills in job performance. *Journal of Applied Psychology,* 1961, **45**(3), 137-142. (b)

Schultz, D., & Siegel, A. *The rationale and application of job suitability as a basis for the evaluation of training.* Personnel Psychology, 1962, **15**(3), 261-278. (a)

Schultz, D., & Siegel, A. *Post-training performance, criterion development and application: A multidimensional scaling analysis of the job performance of Naval aviation electronics technicians.* Wayne, Pa.: Applied Psychological Services, 1962. (b)

Schultz, D., & Siegel, A. *Post-training performance, criterion development and application: A multidimensional scaling analysis of the circuit types repaired by Naval aviation electronics technicians.* Wayne, Pa.: Applied Psychological Services, 1963.

Schultz, D., & Siegel, A. *Post-training performance, criterion development and application: The development of unidimensional scales for the dimensions derived from a multidimensional scale analysis of the job of the Naval aviation electronics technicians.* Wayne, Pa.: Applied Psychological Services, 1964.

Scott, J., & Phelan, J. Expectancies an unemployable makes regarding source of control of reinforcement. *Psychological Reports,* 1969, **25**(3), 911-913.

Scriven, M. The methodology of evaluation. *Perspectives of curriculum evaluation.* Chicago: Rand McNally & Co., 1967. Pp. 39-83.

Seidel, R. *Discussion of a unique approach to CAI: Project Impact. Innovations for training.* HumRRO Professional Paper No. 6-69, 1969, 10-24.

Sheppard, W., & MacDermot, H. Design and evaluation of a programmed course in introductory psychology. *Journal of Applied Behavior Analysis,* 1970, **3**, 5-11.

Showel, M., Taylor, E., & Hood, P. *Automation of a portion of NCO leadership preparation training.* HumRRO Technical Report No. 66-21, 1966.

Shuford, E. *Confidence testing: A new tool for measurement.* Lexington, Mass.: Shuford-Massengill, Corp., 1967.

Shuford, E., Albert, A., & Massengill, H. Admissible probability measurement procedures. *Psychometrika,* 1966, **31**, 125-145.

Siegel, A. *Development of performance evaluative measures. Personnel Psychophysics: Terminal threshold and signal detection theoretic applications to performance assessment.* Wayne, Pa.: Applied Psychological Services, 1968.

Siegel, A., & Federman, P. *Development of a method for deriving required training aids/devices and application to the tactical coordinator position in ASW aircraft.* Wayne, Pa.: Applied Psychological Services, 1969.

Siegel, A., & Federman, P. *Development of performance evaluation measures. Investigation into and application of a fleet post-training performance evaluation system.* Wayne, Pa.: Applied Psychological Services, 1970.

Siegel, A., Federman, P., & Richlin, M. *Post-training performance, criterion development and application. The SESR program: Commissioned and petty officer opinions.* Wayne, Pa.: Applied Psychological Services, 1959.

Siegel, A., & Fischl, M. *Mass training techniques in civil defense: II. A further study of telephone adjunct training.* Wayne, Pa.: Applied Psychological Services, 1965.

Siegel, A., Fischl, M., & Pfeiffer, M. *Personnel psychophysics: Terminal threshold and signal detection theoretic applications to performance assessment.* Wayne, Pa.: Applied Psychological Services, 1968.

Siegel, A., & Pfeiffer, M. *Post-training performance, criterion development and application. Personnel psychophysics: Estimating personnel subsystem reliability through magnitude estimation methods.* Wayne, Pa.: Applied Psychological Services, 1966. (a)

Siegel, A., & Pfeiffer, M. *Post-training performance, criterion development and application. Personnel psychophysics: Operational correlates of electronic circuit complexity.* Wayne, Pa.: Psychological Services, 1966. (b)

Siegel, A., & Pfeiffer, M. Predicting academic success through application of theory of signal detectability variables. *Proceedings of the 77th Annual Convention of the APA,* 1969, 145-146.

Siegel, A., Richlin, M., & Federman, P. *Post-training performance, criterion development and application: Development and application of TBCL criteria to the SESR program for the air controlman and the parachute rigger ratings.* Wayne, Pa.: Applied Psychological Services, 1958.

Siegel, A., & Schultz, D. *Post-training performance, criterion development and application: Generalized Thurstone and Guttman scales for electronic job performance evaluation.* Wayne, Pa.: Applied Psychological Services, 1960.

Siegel, A., & Schultz, D. Evaluating the effects of training. *Journal of American Society of Training Directors,* 1961, (Reprint).

Siegel, A., & Schultz, D. *Post-training performance, criterion development and application: A comparative multidimensional scaling analysis of the tasks performed by Naval aviation electronics technicians at two job levels.* Wayne, Pa.: Applied Psychological Services, 1963.

Siegel, A., Schultz, D., & Benson, S. *Post-training performance, criterion development and application: A further study into technical performance checklist criteria which meet the Thurstone and Guttman scalability requirements.* Wayne, Pa.: Applied Psychological Services, 1960.

Siegel, A., Schultz, D., & Federman, P. *Post-training performance, criterion development and application: A matrix method for the evaluation of training.* Wayne, Pa.: Applied Psychological Services, 1961.

Siegel, A., Schultz, D., & Lanterman, R. *The development and application of absolute scales of electronic performance.* Wayne, Pa.: Applied Psychological Services, 1964.

Sieveking, N., & Larson, G. Analysis of the American Chemical Society achievement test with a multivariate prediction of college chemistry achievement. *Journal of Consulting Psychology,* 1969, **16**(2), 166-171.

Simon, G. Comments on "Implications of criterion-referenced measurement." *Journal of Educational Measurement,* 1969, **6**, 259-260.

Smode, A., Hall, E., & Meyer, D. *An assessment of research relevant to pilot training.* Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratories, 1966.

Staff Members of Division No. 5 (Air Defense). *The effectiveness and implementation of instructional closed-circuit television.* Collected papers prepared under work unit TEXTRUCK. Methods of instruction in technical training. HumRRO Professional Paper No. 34-70, 1970, 17-31.

Stake, R. Generalization of program evaluation: The need for limits. *Educational Product Report,* 1969, **2**, 39-40.

Stake, R., & Denny, T. Needed concepts and techniques for utilizing more fully the potential of evaluation. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 370-390.

Standlee, L., & Hooprich, E. *Review of research on reading instruction for adults.* San Diego: Naval Personnel Research Activity, 1962.

Standlee, L., & Saylor, J. *Second study of equipment operator class 'A' school training. For group IV personnel.* San Diego: Naval Personnel Research Activity, 1969.

Steadman, J., Bilinski, C., Coady, J., & Steinemann, J. *The development and evaluation of training methods for group IV personnel. II. Training group IV personnel in the electronic multimeter AN/PSM-4.* San Diego: Naval Personnel Research Activity, 1969.

Steadman, J., & Harrigan, R. *Evaluation of DS technician graduates of the set six year obligor training program.* Research Report No. SRR-71-18. San Diego: Naval Personnel and Training Research Laboratory, 1971.

Steinemann, J., Coady, J., Harrigan, R., Matlock, E., & Steadman, J. *Evaluation of ET graduates of the set six year obligor training program.* San Diego: Naval Personnel and Training Research Laboratory, 1969.

Steinemann, J., Coady, J., Harrigan, R., & Matlock, E. *Evaluation of graduates of the electronics technician phase A-1 training program.* San Diego, California: Naval Personnel Research Activity, 1968.

Steinemann, J., Hooprich, Archibald, A., & Van Matre, A. *Development of a "wordsmanship" training course for marginal personnel.* Research Report No. SRR-71-17. San Diego, California: Naval Personnel and Training Research Laboratory, 1971.

Stewart. J. *The usefulness of task analysis in the evaluation of military training.* Monterey, California: Navy Postgraduate School, Master's Thesis, 1970.

Sticht, T. *Learning by listening in relation to aptitude, reading, and rate controlled speech.* Technical Report 69-23. HumRRO Division No. 3, 1969.

Stone, L., & Sinnett, E. Academic grades: Their rationale and empirical scale structure. *Psychological Reports,* 1968, **22**(3), 681-686.

Suchman, E. *Evaluative research: Principles and practice in public service and social action programs.* New York: Russell Sage Foundation, 1967.

Systems Approach to Course Development. *Instructional system development course.* Sheppard AFB, ATC Study Guide 3 AIR 75130-X-1, 1970.

Tallmadge, G. Relationship between training methods and learner characteristics. *Journal of Educational Psychology,* 1968, **59**(1), 32-36.

Taylor, J. *Factors related to individual training.* HumRRO Professional Paper No. 11-70, 1970.

Taylor, P. Where angels fear to tread. *Record,* 1970, **71**, 357-369.

Thelen, H. The evaluation of group instruction. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969. Pp. 115-155.

Trooboff, B. *Employment experience after MDTA training: A study of the relationship between trainee characteristics and post-training employment experience.* Atlanta, Georgia: Georgia State College School of Business Administration, 1968.

Trow, M. Methodological problems in the evaluation of innovation. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York; Holt, Rinehart & Winston, 1970. Pp. 289-305.

Tucker, L. Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. Cattell (Ed.), *Handbook of multivariate experimental psychology.* Chicago: Rand McNally, 1966.

Tyler, R. (Ed.) *Educational evaluation: New roles, new means.* Chicago: University of Chicago Press, 1969.

Underwood, B. Some correlates of item repetition in free-recall learning. *Journal of Verbal Learning and Verbal Behavior,* 1969, **8**, 83-94.

Underwood, B. A breakdown of the total-time law in free-recall learning. *Journal of Verbal Learning and Verbal Behavior,* 1970, **9**, 573-580.

Valverde, H. *A systems approach to electronic maintenance training.* Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory, 1969.

Van Matre, N., & Harrigan, R. *A performance oriented electronics technician training program. V. Final fleet follow-up of graduates.* Research Report No. SRR-70-15. San Diego, California: Naval Training Research Laboratory, 1970.

Van Matre, N., & Steinemann, J. *A performance oriented electronics technician training program. II. Initial fleet follow-up evaluation of graduates.* Technical Bulletin STB-67-15. San Diego, California: U.S. Naval Research Laboratory, 1966.

Veldman, D., & Peck, R. Influences on pupil Evaluations of student teachers. *Journal of educational Psychology,* 1969, **60**(2), 103-108.

Vineberg, R., Sticht, T., Taylor, E., & Caylor, J. *Effects of aptitude (AFQT) job experience, and literacy on job performance: Summary of HumRRO work units UTILITY and REALISTIC.* HumRRO Division No. 3, 1970.

Waina, R. *Specification of educational objectives for system evaluation.* Santa Monica, California: Rand Corporation, 1969.

Walker, R. An evaluation of training methods and their characteristics. *Human Factors,* 1965. **7**(4), 347-354.

Ward, J., Fooks, N., Kern, R., & McDonald, R. *Development and evaluation of an integrated basic combat/advanced training program for medical corpsmen.* HumRRO Technical Report No. TR-70-1, 1970.

Watson, P. An industrial evaluation of four strategies of instruction. *Audio visual instruction,* 1968. 156-158.

Weiss, R., & Rein, N. The evaluation of broad aim programs: Experimental design, its difficulties, and an alternative. *Administrative Science Quarterly,* 1970, **15**, 97-109.

Weitz, J. *The use of criterional measures.* Technical Report No. 1, ONR **285**(51), 1962.

Weitz, J. The use of criterional measures. *Psychological Report,* 1964, **14**, 803-817.

Westbrook, B., & Jones, C. The reliability and validity of a class-constructed measure of achievement in tests and measurements. *Educational and Psychological Measurement,* 1968, **28**(2), 485-486.

Whitla, D. Research in college admissions. In R. Tyler (Ed.), *Educational evaluation: New roles, new means.* Chicago: University of Chicago, Press, 1969, Pp. 81-101.

Whitmore, P. *A rational analysis of the process of instruction.* Collected papers prepared under work unit TEXTRUCK. Methods of instruction in technical training. HumRRO Professional Paper No. 34-70, 1970, 51-68. (a)

Whitmore, P. *Deriving and specifying instructional objectives.* Collected papers prepared under work unit TEXTRUCK. Methods of instruction in technical training. HumRRO Professional Paper No. 34-70, 1970, 38-47. (b)

Whitmore, P. *Automated instructional methods for technical training.* Collected papers prepared under work unit TEXTRUCK. Methods of instruction in technical training. HumRRO Professional Paper No. 34-70, 1970, 32-37. (c)

Whitmore, P. *Developing new instructional techniques.* Collected papers prepared under work unit TEXTRUCK. Methods of instruction in technical training. HumRRO Professional Paper No. 34-70, 13-16. (d)

Whitmore, P., Cox, J., & Friel, D. A classroom method of training aircraft recognition. HumRRO Technical Report No. 68-1, 1968.

Wiley, D. Design and analysis of evaluation studies. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction.* New York, Holt, Rinehart & Winston, 1970. Pp. 259-269.

Winch, R., & Campbell, D. Proof? No. Evidence? Yes. The significance of tests of significance. *American Sociologist,* 1969, 140-143.

Wittrock, M. The evaluation of instruction: Cause-and-effect relations in naturalistic data. In M. Wittrock & D. Wiley (Eds.) *The evaluation of instruction.* New York: Holt Rinehart & Winston, 1970, Pp. 3-21.

Wittrock, M. & Wiley, D. (Eds.) *The evaluation of instruction.* New York: Holt, Rinehart & Winston, 1970.

Wood, D. *Test construction.* Columbus, Ohio: Merrill, 1960.

Worchel, S., & Mitchell, T. *An evaluation of the effectiveness of the culture assimilator in Thailand and Greece.* Report No. TR-70-13. Seattle, Washington: Washington University, Department of Psychology, 1970.

Yeager, J., & Kissel, M. *An investigation of the relationship between selected student characteristics and time required to achieve unit mastery.* Learning Research and Development Center: University of Pittsburgh, 1969.

Yellen, T. *Validation of the MOS evaluation test for field artillery crewman. MOS code 13B40 and the enlisted efficiency report.* Indianapolis, Indiana: Army Enlisted Evaluation Center, 1969.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Applied Psychological Services, Inc.
Wayne, Pennsylvania

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
TRAINING EVALUATION AND STUDENT ACHIEVEMENT MEASUREMENT: A REVIEW OF THE LITERATURE

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(First name, middle initial, last name)*
Brian A. Bergman
Arthur I. Siegel

6. REPORT DATE
January 1972

7a. TOTAL NO. OF PAGES
57

7b. NO. OF REFS
316

8a. CONTRACT OR GRANT NO.
F41609-71-C-0025

b. PROJECT NO 1121

c. Task No. 112103

d. Work Unit No. 112103004

9a. ORIGINATOR'S REPORT NUMBER(S)
AFHRL-TR-72-3

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT
Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Air Force Human Resources Laboratory
Technical Training Division
Lowry Air Force Base, Colorado 80230

13. ABSTRACT

The current training evaluation and student measurement literature is reviewed. The emphasis is on studies which have been reported in the last ten years, although earlier studies which have impacted heavily on recent trends are also included. Because of the obvious interaction between both training evaluation and student measurement, on the one hand, and such topics as statistical methods, methods for course development, training methods, learning styles, motivation, and moderator variables, on the other hand, these and similar considerations are also included.

DD FORM 1 NOV 65 1473

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
technical training						
education						
testing						
course evaluation						
student measurement						
quantitative methods						
training analysis						
training analysis						
training evaluation						
student achievement						
individual differences						